

# Context-aware Sentiment Analysis: A Lexicon-based Machine Learning Approach

DISSERTATION

zur Erlangung des akademischen Grades

**Doktor der technischen Wissenschaften**

eingereicht von

**Stefan Gindl**

Matrikelnummer 9925024

an der  
Fakultät für Informatik der Technischen Universität Wien

Betreuung: ao. Univ.-Prof. Dr. Dieter Merkl, ao. Univ.-Prof. DDr. Arno Scharl

Diese Dissertation haben begutachtet:

---

(ao. Univ.-Prof. Dr. Dieter  
Merkl)

---

(ao. Univ.-Prof. DDr. Arno  
Scharl)

Wien, 09.02.2015

---

(Stefan Gindl)



# Context-aware Sentiment Analysis: A Lexicon-based Machine Learning Approach

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

**Doktor der technischen Wissenschaften**

by

**Stefan Gindl**

Registration Number 9925024

to the Faculty of Informatics  
at the Vienna University of Technology

Advisor: ao. Univ.-Prof. Dr. Dieter Merkl, ao. Univ.-Prof. DDr. Arno Scharl

The dissertation has been reviewed by:

---

(ao. Univ.-Prof. Dr. Dieter  
Merkl)

---

(ao. Univ.-Prof. DDr. Arno  
Scharl)

Vienna, 09.02.2015

---

(Stefan Gindl)



# Erklärung zur Verfassung der Arbeit

Stefan Gindl  
Weißwolffgasse 12, 1210 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 Verfasser)



# Acknowledgements

I am deeply thankful to all the people who made this thesis possible. Their inspiration and support helped me overcome the many obstacles I faced during my work. First of all, I would like to thank my supervisors Arno Scharl and Dieter Merkl. Arno Scharl was my second supervisor and project leader. He helped me plan the work for my thesis and made me rethink many assumptions. Furthermore, he tirelessly helped me iron out flaws and inadequacies in my argumentation. His constant strive for perfection was a great motivation for me. I also want to heartfully thank my first supervisor, Dieter Merkl. He made me trust in myself and gave me the strength to pursue my work. He critically reflected on my statements and made me question arguments I had taken for granted. I also want to thank Albert Weichselbraun, who greatly inspired my work and helped me with the implementation of the prototype. He gave my thesis a sound technical and mathematical background and was always aware of the most efficient technologies to be used in my prototype. I am also thankful for my colleagues at MODUL University Vienna. I had a great time working with them and had many insightful conversations, which inspired my work significantly.

Most importantly, I want to thank my girlfriend Anja, who was a source of strength whenever mine was low. She helped me to stay focused and to keep my objective in view.

Finally, I want to thank my mother. Without you, this would not have been possible.





# Abstract

Sentiment analysis, the research area focusing on the creation, implementation, and evaluation of systems for the analysis of human attitudes, has become increasingly interesting for researchers of diverse special fields such as artificial intelligence, computational linguistics, or psychology. With the wide availability of opinionated statements on the Web and the creation of ever more powerful algorithms, the research area has gotten off the sidelines and moved into the focal point of many scientific projects. It has a significant business value, as it is a central component of media intelligence systems, supporting decisions for marketing campaigns and collecting customer feedback from the large pool of opinions on the Web. It helps decision makers to understand trends on the market, which eventually helps to adapt current marketing strategies. Sentiment analysis also proves beneficial in the political area, by evaluating a political campaign or to measure public awareness towards events of public interest, e.g. climate change or wars. An elicitation of opinions on such a large scale was inconceivable in the era before the World Wide Web and becomes feasible merely because of the existence of powerful technologies, such as machine learning and natural language processing.

This work aims at improving a central resource crucial in sentiment analysis, the sentiment lexicon. These collections of opinionated terms store a-priori charges for each term, indicating whether a term conveys positive or negative sentiment. The charges are bound to manual assessment, even in cases where a term is ambiguous and might change its charge depending on the context. For instance, the term “cool” triggers opposite emotions in the sentence “the cool car” and “she mustered him with a cool glance”. These polarity changes limit approaches which depend on static a-priori charges. The present work expands the sentiment lexicon with context terms, i.e. terms frequently co-occurring with the sentiment term. Analysing their frequency of co-occurrence in positive and negative contexts and storing the probability of co-occurrence results in the creation of *contextualized lexicons*. The probabilities for positive and negative context supersede the fixed a-priori values. A system armed with such a tool is capable of flexibly adapting the sentiment value of one and the same term based on the context it is used in.

A formal evaluation shows the efficacy of the approach. The evaluation follows a method well-established in sentiment analysis: a corpus consisting of product and service reviews from different domains is the basis for the evaluation. Calculating recall, precision, and f-measure in a ten-fold cross-validation shows that the proposed approach outperforms a traditional keyword lookup algorithm with fixed polarities.



# Kurzfassung

Sentimentanalyse ist jenes Forschungsfeld, dass sich mit der Konzeption, Implementierung und Evaluierung von Systemen beschäftigt, die menschliche Stimmungen verstehen sollen. Durch die breite Verfügbarkeit von stimmungsgeladenen Aussagen im World-Wide-Web und leistungsstarken Algorithmen zu deren Analyse, hat sich das Forschungsfeld von seinem Nischendasein zu einem zentralen Bestandteil vieler Forschungsprojekte entwickelt. Sein hoher wirtschaftlicher Wert ergibt sich aus seiner zentralen Rolle in Media-Intelligence-Systemen. Diese unterstützen Marketing-Kampagnen und sammeln KundInnenfeedback aus dem großen Pool von online verfügbaren, geschriebenen Meinungen. Entscheidungsträger können dadurch aktuelle Markttrends leichter nachvollziehen und Marketingstrategien dementsprechend anpassen. Sentimentanalyse erweist sich auch im politischen Bereich als nützliches Werkzeug. Politische Kampagnen lassen sich damit evaluieren und sie unterstützt dabei, die Stimmung bei Ereignissen von öffentlichem Interesse zu messen, etwa dem Klimawandel oder einem Krieg. Meinungsforschung wird dadurch in einem so großen Stil möglich, wie sie vor Zeiten des World-Wid-Web undenkbar gewesen wäre. Die Verfügbarkeit leistungsstarker Rechner gestattet es, komplexe Algorithmen, etwa aus dem maschinellen Lernen oder der natürlichen Sprachverarbeitung, in angemessener Zeit auszuführen.

Die vorliegende Arbeit beschäftigt sich damit, eine zentrale Ressource der Sentimentanalyse zu verbessern: das Sentimentlexikon. Dieses Lexikon enthält stimmungstragende Terme zusammen mit einer Einschätzung ihrer Polarität. Diese Stimmungsladung wird händisch ermittelt und ist statisch, selbst in Fällen, wo sich die Ladung eines Wortes durch den Kontext, in dem es verwendet wird, verändern kann. Das Wort "kühl" löst beispielsweise in "ein kühler Kopf" oder "ein kühler Blick" entgegengesetzte Empfindungen aus. Diese Ladungsveränderungen limitieren die Leistungsfähigkeit von Systemen, die von statischen Ladungen ausgehen. In der vorliegenden Arbeit werden Sentimentlexikons mit Kontexttermen erweitert, d.h. mit Termen, die häufig gemeinsam mit bestimmten Sentimenttermen vorkommen. Die Wahrscheinlichkeit des gemeinsamen Auftretens wird im Lexikon mitgespeichert, wodurch aus dem klassischen Sentimentlexikon ein *kontextualisiertes Lexikon* entsteht. Mit so einem Werkzeug ausgestattet ist ein Sentimentanalysesystem in der Lage, sich flexibel an unterschiedliche Kontexte anzupassen.

Eine formelle Evaluierung zeigte die Wirksamkeit des vorgestellten Ansatzes. Sie folgte dabei einer in der Sentimentanalyse üblichen Vorgehensweise, bei der Produkt- und Servicekriterien aus unterschiedlichen Domänen als Evaluierungskorpus herangezogen werden. Durch die Berechnung von Recall, Precision und F-Measure in einer zehnfachen Kreuzvalidierung konnte gezeigt werden, dass der vorgestellte Ansatz einen Schlagwortansatz mit statischen Ladungen übertrifft.



# Contents

<b>List of Tables</b>	<b>xi</b>
<b>List of Figures</b>	<b>xii</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Research Contribution . . . . .	5
1.2 Purpose . . . . .	6
1.3 Structure of the Thesis . . . . .	6
<b>2 Related Work</b>	<b>7</b>
2.1 Sentiment Lexicons . . . . .	8
2.2 Improving Sentiment Lexicons . . . . .	9
2.3 Polarity Classification . . . . .	16
2.4 Opinion Target Extraction . . . . .	24
2.5 Affect Analysis . . . . .	34
2.6 Invocation of Context Information . . . . .	37
<b>3 Description of the Scientific Framework</b>	<b>41</b>
3.1 Design Science Research . . . . .	42
3.2 Research Guidelines in Design Science . . . . .	43
<b>4 Methodology</b>	<b>51</b>
4.1 Legitimation . . . . .	52
4.2 Term Definition . . . . .	56
4.3 Overview . . . . .	57
4.4 Contextualization in Detail . . . . .	60
4.5 Preprocessing . . . . .	65
4.6 Text Classification . . . . .	67
<b>5 Evaluation</b>	<b>75</b>
5.1 The Baseline . . . . .	75
5.2 Efficacy Measurements in Information Retrieval . . . . .	76
5.3 Cross-validation . . . . .	78
5.4 The Evaluation Datasets . . . . .	79

5.5	Results . . . . .	85
5.6	Discussion . . . . .	88
<b>6</b>	<b>Conclusion</b>	<b>91</b>
6.1	Summary . . . . .	91
6.2	Discussion . . . . .	92
6.3	Future Work . . . . .	92
	<b>Bibliography</b>	<b>95</b>

# List of Tables

2.1	Comparison of the initial lexicon with the expanded lexicon. . . . .	15
2.2	Comparison of the expanded lexicon with the General Inquirer. . . . .	16
2.3	Examples of terms added after bootstrapping. . . . .	17
2.4	The three labels used for expression-level annotation. . . . .	22
2.5	Regular expression patterns for opinion aspect extraction. . . . .	32
2.6	The top 15 most frequent positive and negative targets with their respective frequency counts. . . . .	32
2.7	Top 15 strongest positive and negative aspects. . . . .	33
2.8	Example sentences with targets and aspects. . . . .	33
4.1	Example entries in a sentiment lexicon. . . . .	52
4.2	Expanding a sentiment lexicon with POS tags in Penn Treebank style (Marcus et al., 1993). . . . .	53
4.3	Example entries in a contextualized lexicon. . . . .	65
4.4	English negation triggers. . . . .	66
4.5	Overview of the Penn Treebank tag-set. . . . .	67
4.6	Example terms in the used sentiment lexicon. . . . .	73
5.1	Results for TripAdvisor; an arrow indicates a significant gain/loss, dots indicate stagnation. . . . .	85
5.2	Results for Amazon; an arrow indicates a significant gain/loss, dots indicate stagnation. . . . .	86
5.3	Results for IMDb; an arrow indicates a significant gain/loss, dots indicate stagnation. . . . .	87
5.4	Number of total versus significant performance gains and losses. . . . .	88
5.5	Performance gains and losses per review category and measurement. . . . .	89
5.6	Examples for successful disambiguation (ambiguous terms are in bold face, context terms in italics). . . . .	90

# List of Figures

2.1	The Sentiment Quiz was deployed as a Facebook application to increase user engagement. . . . .	12
2.2	Overview of the bootstrapping procedure. . . . .	13
2.3	The evaluation compares the three lexicons using keyword lookup, Naïve Bayes and an SVM classifier. . . . .	15
2.4	The three annotation layers on the example sentence “Journalists should continue to exercise their right to freedom of expression without attacks.” . . . . .	21
2.5	Example sentence with a single opinion holder and target. . . . .	23
2.6	Multiple holders and targets make the annotation complex. . . . .	23
2.7	Opinion Observer neatly contrasts positive/negative sentiment towards aspects of digital cameras. . . . .	27
2.8	<i>The phone has a good screen</i> - <i>screen</i> receives a positive charge from <i>good</i> . . . . .	29
2.9	<i>The phone has a good screen but a bad battery</i> - multiple targets with differing polarities in one sentence. . . . .	30
2.10	Identifying sentiment targets across two sentences. . . . .	31
2.11	The wheel of emotions knows eight basic emotions. . . . .	35
2.12	The hourglass of emotions. . . . .	36
2.13	Visualization of a Skype conversation with Synesketch. . . . .	37
3.1	Design activities in an organizational environment (Henderson and Venkatraman, 1993). . . . .	43
3.2	The framework of design science research with its three pillars. . . . .	44
4.1	Representing ambiguities of polarities in SentiWordNet. . . . .	56
4.2	The contextualization procedure starting with 1) the identification of ambiguous terms, 2) continuing with the creation of the contextualized lexicon and 3) applying it to an unknown document. . . . .	59
4.3	Exemplary frequency distributions of ambiguous and monosemous terms in positive and negative documents. . . . .	61
4.4	Exemplary frequency distributions of ambiguous and monosemous terms in reviews with ratings from one to five stars (three stars excluded). . . . .	61
4.5	Identification of ambiguous terms in a review corpus. . . . .	62
4.6	The contextualization procedure. . . . .	64



4.7	A hyper-plane separates the data points (Manning et al., 2009). . . . .	70
4.8	Different $k$ values result in differing classifications, requiring knowledge about the data to choose the most promising value. . . . .	72
5.1	The test partition moves until every part of the document collection has served as test partition at least once. . . . .	79
5.2	A single review on Amazon. . . . .	82
5.3	Review statistics on Amazon. . . . .	82
5.4	The most helpful positive and negative review of a product. . . . .	83
5.5	A single review on TripAdvisor. . . . .	83
5.6	Review statistics on TripAdvisor. . . . .	83
5.7	A single review on IMDb. . . . .	84
5.8	Review statistics on IMDb. . . . .	84
5.9	Cross-validation on the TripAdvisor data. . . . .	85
5.10	Cross-validation on the Amazon data. . . . .	86
5.11	Cross-validation on the IMDb data. . . . .	87



*Emotion is a sum totaled by an adding machine of the mind.*  
Ayn Rand, *Atlas Shrugged*.



# Introduction

*Before everything else, getting ready is the secret to success.*

Henry Ford

Human communication serves for the exchange of factual knowledge and emotional states. Frequently, these two types are intermixed, resulting in a subjective interpretation of an objective fact. This results in expressions about the personal satisfaction with current environmental characteristics, e.g. the political situation, one's own health status or mundane things such as the weather. The expression of emotion requires two of the human senses: the sense of hearing, either by exchanging emotion-bearing words or subtle tonal changes indicating joy, fear, or anger, and visual perception, either by analyzing the gestures or facial expressions of the dialog partner, or by reading a written text. The latter channel is subject of this thesis. Sentiment analysis, the computer-driven interpretation of written text aims at understanding the emotional state of persons while they were writing the text. A-priori knowledge, such as knowing which words of a language tend to express positive or negative sentiment, contributes the basic pieces and is completed with algorithms for the analysis of grammar, i.e. part-of-speech taggers and parsers, as well as hidden connections, uncovered by machine learning algorithms.

Understanding human emotion is difficult for computers. Human language has many subtleties, e.g. stylistic means such as irony or sarcasm, and requires extensive knowledge about the world that the computer does not possess. For instance, why should it be a bad idea to put your hand closely above a candle if you cannot feel any pain? This lack of common sense makes it hard for computers to interpret emotions in-depth. Furthermore, expressions of human emotions are non-verbal frequently. Facial expressions and changes in the voice of the speakers help the listeners to interpret the emotional state of their dialog partners. Even with all this information at hand it is often tricky to understand the point-of-view of a person, which leads to frequent misunderstanding. Understandably, the lack of this layer of information makes it even harder for

computers to correctly interpret emotions. Thus, sentiment analysis has a natural limitation in how accurate it can be. These limitations make sentiment analysis a challenging, yet fascinating research area.

Sentiment analysis has evolved into a highly attractive research area. The emergence of the Web allowed people to publish and share their opinions online. People can comment on various affairs of their daily life, e.g. political attitudes, the quality of a product just purchased or a holiday trip. Publishing is possible on different web-based media. Micro-blogging services such as Twitter ([www.twitter.com](http://www.twitter.com)), Tumblr ([www.tumblr.com](http://www.tumblr.com)) or Soup ([www.soup.io](http://www.soup.io)) allow for a fast way to share thoughts on various topics, distribute links of interesting websites or upload pictures and videos. Social networking platforms from Google+ ([plus.google.com](http://plus.google.com)) over LinkedIn ([www.linkedin.com](http://www.linkedin.com)) to Facebook ([www.facebook.com](http://www.facebook.com)) radically changed the way people connect with each-other. These platforms are places where opinions and attitudes are shared at a large scale. Another online medium to share information and opinions are forums and blogs. Forums are often specialized on a certain topic and allow for a detailed discussion on aspects of that topic. In a forum, people can also answer questions or provide a statement to clarify or question the statements of other people. Exploiting forums can be valuable for companies, since they represent a collection of opinions on a certain topic. Moreover, thanks to their liveliness, they can also give insight on the variation of opinions, e.g. when people disagreeing with statements offer a new point of view. In contrast, blogs, are strongly individualized media and serve as a channel to express individual views.

All mentioned media contain potential business value for a company. Companies are interested in the acceptance of a product or service, or in the general perception of the company in the public. Telephone surveys or personal questionnaires have traditionally been the only tool to assess people's opinions. Yet, these methods have clear disadvantages, as their accomplishment takes time and requires motivated interviewees. Manually sifting through the opinions available on the Web is another option. However, this strategy requires extensive human labor, which makes it expensive and time-consuming as well. Reading the opinionated texts fonders on the sheer mass of data.

These problems call for a solution, e.g. an automated way of extracting already existing opinions from the Web on a large scale, without the burden of extensive human labor. Sentiment analysis, or opinion mining, is the research area providing the tools to accomplish exactly this. It explores methods from a variety of research areas, such as artificial intelligence, natural language processing, linguistics, or web mining and combines them in a way to extract opinions, expressed by humans on a diverse set of topics. Sentiment analysis investigates unstructured data, i.e. free text, which is, in contrast to structured data, hard to query. The information is weaved into a data type that is still hard to understand for computers: the human language. It suffers from a variety of flaws, e.g. language subtleties, word ambiguities, or terms inter-related in a way that sophisticated algorithms need to be employed to unravel the connections. Furthermore, humans weave an extensive knowledge about the world into their language, adding another layer of difficulty. Metaphors, irony, or sarcasm, in some cases even unintelligible for other humans, become entirely cryptic for the computer.

Sentiment analysis aims at tackling these problems. Sophisticated machine learners combined with powerful natural language processing algorithms provide the basis for a linguistic

understanding of language, and linguistic rules crack difficult language structures. The development of these tools within the last years has made it possible to analyze opinions available in mankind's most extensive information store, the World Wide Web. Sentiment analysis provides the methods to overcome the mentioned problems. It does not require to create internal incentives for people to express their opinions - the opinions are already there. Since people voluntarily publish them on the web it is not necessary to spend time and money to express them explicitly. Furthermore, sentiment analysis allows for an analysis of extensive document collections. Manually reading through a large number of documents fails due to the limitations humans have: reading, understanding and analyzing text requires a lot of time and humans tend to lose concentration after a while, resulting in poor quality of labor.

## 1.1 Research Contribution

One challenge in sentiment analysis is to determine the polarity of a term, i.e. whether it expresses positive or negative sentiment. Knowing the sentiment of a term helps to construct so-called *sentiment lexicons*, i.e. collections of sentiment-bearing terms. They serve as the basic pillar for a variety of algorithms by delivering a-priori sentiment knowledge.

The compilation of a sentiment lexicon is challenging: firstly, it can be hard to decide whether a term expresses sentiment or not; secondly, it is hard to determine the sentiment strength, i.e. is it a *very* negative term or just a negative term; thirdly, sentiment terms can switch their polarity in special circumstances. For instance, the term *killer* usually denotes a person doing harm to other people, invoking negative emotion. However, when used in the sentence *his new novel is a killer*<sup>1</sup>, the same term conveys positive sentiment (section 4.1 contains an elaborate description of this matter of fact). This latter problem, i.e. the identification of polarity switches triggered by differences in context, is the topic of this thesis.

During this research we created an artifact that adds a new layer of accuracy to sentiment lexicons by turning them into so-called *contextualized lexicons*. The artifact separates ambiguous from monosemous sentiment terms in a sentiment lexicon via corpus analysis. The corpus consists of a set of positive and negative documents. Ambiguous terms are terms whose frequency in a labeled corpus indicates that they are used in positive and negative texts alike. Subsequently, the artifact extracts context terms for each ambiguous term, i.e. terms frequently co-occurring with the ambiguous term. By employing the Naïve Bayes technique the artifact stores probability values for each ambiguous term and its context terms, indicating whether an ambiguous term/context term co-occurrence is more likely in a positive or a negative document. This additional layer of information turns a traditional sentiment lexicon into a contextualized lexicon, which becomes a valuable resource in every sentiment analysis system.

---

<sup>1</sup>The Oxford Dictionaries, <http://www.oxforddictionaries.com/definition/english/killer>, last accessed on 24 November 2014

## 1.2 Purpose

The purpose of sentiment analysis is manifold. From a scientific perspective it gained interest in areas such as artificial intelligence, computational linguistics, or psychology. The desire to understand the human brain, although still far beyond the reach of modern computer technology, has challenged researchers across the world to seek methods to imitate it. Employing them in next-generation robots or intelligent agents is an ultimate goal.

From a business perspective, sentiment analysis has turned from a nice-to-have to a must-have in the portfolio of a company's marketing tools. Launching a marketing strategy and monitoring it via media monitoring services is now in the standard repertoire of leading companies. Neglecting this kind of information invokes the risk of missing unwanted media attention, such as flame-wars on social media platforms. From a political perspective, sentiment analysis helps to monitor political campaigns, to discover flaws in the argumentation of politicians invoking negative response, or to measure the reaction of the public towards events such as strikes. Monitoring them helps to adapt accordingly, avoiding that negative sentiment piles up until it bursts.

Solutions for sentiment analysis are commercially available, and differ in their levels of maturity. The services offered range from news media and social media analysis over social TV show analysis to support for publishing content online. Professional tools such as webLyzard, ([weblyzard.com](http://weblyzard.com)), Luminoso ([luminoso.com](http://luminoso.com)), Netbase ([netbase.com](http://netbase.com)), Attensity ([attensity.com](http://attensity.com)), Radian6 ([www.salesforcemarketingcloud.com](http://www.salesforcemarketingcloud.com)), Converseon ([converseon.com](http://converseon.com)), or TheySay ([theysay.io](http://theysay.io)) offer mature toolkits with sentiment analysis as an essential part of their pipeline. Despite their maturity, these system still benefit from leveraging common knowledge and a deeper knowledge about the world, allowing them to reason on a higher level. This thesis aims to contribute at exactly this point, by adding contextual knowledge to existing approaches, resulting in a better understanding of the true meaning of an opinionated statement.

## 1.3 Structure of the Thesis

An overview of the state-of-the-art in this research area as well as approaches related to the one pursued in this work is the starting point of this thesis (see chapter 2). The description of the scientific framework, i.e. design science research, follows this overview in chapter 3. The chapter also summarizes seven guidelines that were fundamental during the implementation and evaluation procedure in this thesis. Subsequently, the description of the approach in chapter 4 introduces into the theoretical and practical concepts behind this approach. An extensive evaluation in chapter 5 emphasizes the efficacy of the approach. The evaluation leverages an approach well-known in the literature, by using a collection of product and service reviews as an already annotated resource. The conclusion in chapter 6 summarizes the thesis and gives insight into potential paths for future work.



## Related Work

*We have always been shameless about stealing great ideas.*

Steve Jobs, *Triumph of the Nerds*.

Early work on sentiment analysis started with the identification of subjective sentences (Wiebe, 1994) or the discrimination of positive and negative adjectives by exploiting the mutual information of known sentiment indicators and unknown adjectives by analyzing their syntactical relations (Hatzivassiloglou and McKeown, 1997). Building upon these foundations, sentiment analysis is now a well-researched area, with differentiated approaches. The reason for its attractiveness clearly lies in the easy availability of opinions on the Web. A manual analysis, i.e. humans browsing through the Web in search for opinionated statements, is beyond what is feasible due to the sheer amount of available data. Automated techniques step in here, providing the means to funnel the information relevant for the interested observer. The applied techniques can be roughly divided into two main approaches: lexical approaches, which usually employ a kind of sentiment lexicon (i.e. a collection of terms conveying sentiment), and machine learning techniques. The distinction into two main areas is not totally correct though, as most approaches are combinations of the two. The usage of sentiment lexicons is ubiquitous though, hardly any approach relinquishes a sentiment lexicon. Early work used small sets of lexicons (Turney, 2001) and research partially focused on the creation of the necessary resources. As already mentioned, sentiment lexicons are lists of known sentiment terms, where each term has a *sentiment value* (mostly a numerical value ranging from  $[-1, 1]$ ) assigned to it. For instance, *excellent* conveys positive sentiment, which results in the assignment of the positive value 1. *Terror*, commonly indicating something negative, could have the negative value -1. The assignment of sentiment values is a skillful task and requires expert knowledge. Differentiation between subjective and objective terms is necessary, i.e. categorizing them into those terms that convey sentiment and those that do not. This task seems to be straight-forward, but on a closer look it reveals its

difficulty. Some terms convey sentiment quite strongly, such as *beautiful*, *terrific*, or *horrible*. Others are more subtle, e.g. *pure*, *abort*, or *weak*. Then again, others depend on context: they have a positive meaning in one context and a negative in the other. For instance, being *addicted to heroin* is undesirable, while being *addicted to Beethoven's moonlight sonata* expresses joy and positive attitude towards the composer. These subtleties in language make sentiment analysis a challenging task and call for methods to handle them. The presented research work aims at tackling one of these challenges, i.e. the problem of context and how to manage it. The research follows the principles of design science (Hevner et al., 2004).

In the following, we provide an overview of opinion mining approaches and illuminate their application areas and used techniques. We attempt to classify them into the top-level branches “polarity classification”, “opinion holder and target detection”, and “affect analysis”. The first branch, polarity classification, covers the categorization of documents, sentences, phrases, or words into positive and negative. The second branch, opinion holder and target detection, answers the question “who thinks what about whom?”. The last branch, affect analysis, operates on a fine-grained emotional level. It gives the exact orientation of the expressed sentiment, e.g. “rage” or “anger”.

Similar to the classification into lexical approaches and machine learning approaches, the research areas strongly overlap, e.g. modeling an affect analysis task might still result in a summary on whether a text snippet is substantially positive or negative, turning it into a polarity classification task. The overview does not attempt to be complete, as sentiment analysis is a wide and fast-growing research area. For further information, the works by Liu (2012) and Medhat et al. (2014) serve as a good starting point.

## 2.1 Sentiment Lexicons

Sentiment lexicons are the core component of a sentiment analysis system. They are collections of terms with a polarity label attached. For English there is a viable choice of extensive lexicons. In other languages, resources are still sparse.

### Sentiment Lexicons for English

**General Inquirer:** this widely known linguistic resource is not specifically targeted at sentiment analysis. It comprises 182 categories in total, including classifications such as “social”, “food”, or “travel”. The categories most relevant for sentiment analysis are “positive”, “negative”, “strong”, “weak”, and more generic categories such as “pleasure”, “pain”, “feel”, or “arousal” (Stone, 1966). With its broad coverage the General Inquirer is a highly valuable resource.

**Opinion Lexicon:** a lexicon containing approximately 6 800 sentiment bearing words, first used by Hu and Liu (2004) to extract product features discussed in reviews.

**Subjectivity Lexicon:** a list of 8 000 sentiment terms (Wilson et al., 2005). It combines the General Inquirer with a resource created by Hatzivassiloglou and McKeown (1997) and Riloff and Wiebe (2003) and is completed by manual annotation.

**SentiWordNet:** based on WordNet, this resource stores the degree of negativity, positivity, and objectivity (Baccianella et al., 2010; Esuli and Sebastiani, 2006).

**WordNet-Affect:** this resource covers 2 874 WordNet synsets, resulting in 4 787 words in total.

**SenticNet:** SenticNet combines the knowledge of several other knowledge bases. It derives common knowledge, i.e. factual or lexical knowledge of the world, such as “The sun is a star”, from DBPedia Lehmann et al. (2014), WordNet Fellbaum (1998) and Probase Wu et al. (2012), and common-sense knowledge, i.e. knowledge about how the world works, such as “If I touch the burning candle I will feel a sense of pain in my hand”, from ConceptNet Speer and Havasi (2013)

## Other Languages

**German Polarity Clues:** this semi-automatically constructed resource contains 10 141 sentiment terms for German and is publicly available (Waltinger, 2010).

**SentimentWortSchatz (SentiWS):** this publicly available German resource contains 1 650 negative and 1 818 positive words, as well as their inflections.

**Czech Sentiment Lexicon:** Veselovská et al. (2014) created this resource via machine translation of the subjectivity lexicon (Wilson et al., 2005) and a subsequent manual refinement. It contains 1 672 positive and 2 863 negative terms.

**HowNet Sentiment Lexicon:** this common-sense knowledge base Dong et al. (2010) also contains 3 730 positive and 3 116 negative words.

## 2.2 Improving Sentiment Lexicons

Turney and Littman (2002) examine two different word association measurements for learning the polarity of unknown terms. The association measurements point-wise mutual information (PMI; Church and Hanks (1989)) and latent semantic analysis (LSA; Landauer and Dutnais (1997)) give information on the relatedness of two terms. The authors postulate that unknown terms with a strong relatedness to positive sentiment terms also carry positive polarity and vice versa. To prove this hypothesis, they generated a list with seed terms, which they call their paradigm terms, since they are judged to be very secure sentiment terms. Positive paradigm terms are *good, nice, excellent, positive, fortunate, correct, and superior*, the negative paradigm terms are *bad, nasty, poor, negative, unfortunate, wrong, and inferior*. Their system then identifies new strongly related terms from a very large corpus for PMI, containing approximately 100 billion terms, and a smaller corpus for both PMI and LSA, containing roughly ten million terms. The larger corpus is nothing else than all web pages indexed by the search engine AltaVista. AltaVista provides a NEAR operator, returning documents where two query terms must occur in a certain spatial distance. For LSA the usage of a smaller corpus is more feasible, since LSA is a more computationally expensive technique than PMI. The evaluation is accomplished by comparing the polarity of the new terms with the polarity of terms in the General Inquirer.

The results show that PMI applied to the large corpus can compete with the method proposed by Hatzivassiloglou and McKeown (1997). Comparison of PMI and LSA on the smaller corpus shows the superiority of LSA over the simpler PMI technique. This outcome is inverse to the outcome of (Turney, 2001). Here, the author uses PMI and LSA to identify synonyms of terms used in the TOEFL test. With this method a comparison between the automatically extracted synonyms and the manually compiled synonyms of the TOEFL test is possible. In this work PMI clearly outperforms LSA. A potential explanation is that in sentiment detection synonymy is not a good indicator for polarity similarity. It is possible that sentiment is inherent in a more complex and subtle way, which would explain LSA's superiority.

Turney and Littman (2003) further compare point-wise mutual information and latent semantic analysis. A paradigmatic term list, i.e. a list containing undoubted sentiment terms, such as *good* or *bad*, is the starting point for the lexicon expansion. Both previously mentioned measurements collect terms associated to the paradigm terms from three Web page corpora. To evaluate the quality of the procedures, the authors compare the extracted terms with the sentiment lexicons of Hatzivassiloglou and McKeown (1997) and Stone (1966). According to the results of this works LSA outperforms the simple PMI technique.

Neviarouskaya et al. (2009) show techniques for the expansion of a sentiment lexicon. Their lexicon, *SentiFul*, origins from the Affect database (Neviarouskaya et al., 2007). The Affect database contains approximately 2 438 sentiment terms, divided into nine emotive categories. The authors considered three of these as being mainly positive (*interest*, *joy*, and *surprise*) and six as negative (*anger*, *disgust*, *fear*, *guilt*, *sadness*, and *shame*). Each term has an intensity score, ranging from 0 to 1 (e.g. *tremendous* has the intensity 1 in the category *surprise*, 0.5 in category *joy*, and 0.1 for *fear*). The ratio of all positive/negative intensities to the number of positive/negative classes the term occurs in is their polarity score. By dividing the number of a term's positive/negative categories with the number of all its categories another score is calculated, called the polarity weight. The lexicon resulting from these procedures serves as the basis for expansion. The first expansion attempt comprised the exploitation of SentiWordNet (Esuli and Sebastiani, 2006), using two different techniques. The first technique uses only the first SentiWordNet synset to obtain a sentiment value for a term. The other calculates averages over all synsets one and the same term belongs to. Their next attempts comprised the usage of WordNet and a syntactically inspired approach. The WordNet approach exploited direct synonyms of SentiFul terms. For each term in SentiFul related synsets are retrieved. The average sentiment value of SentiFul terms already contained in that synset serve as sentiment values for new terms. Duplicates obtained by this step are eliminated by again assigning the average sentiment value to the term occurring in duplicates. This process expands SentiFul with approximately 4 000 new terms. The authors also pursue a more linguistically inspired approach, extending the database by SentiFul terms with certain affixes attached. Affixes can be either prefixes, attached at the beginning, or suffixes, attached at the end of a term. They discriminate four types of affixes: (1) *propagating* (e.g. 'en' + 'rich'  $\Rightarrow$  'enrich'), (2) *reversing* (e.g. 'harm' + 'less'  $\Rightarrow$  'harmless'), (3) *intensifying* (e.g. 'super' + 'hero'  $\Rightarrow$  'superhero'), and (4) *weakening affixes* (e.g. 'semi' + 'sweet'  $\Rightarrow$  'semisweet'). This process also includes approximately 4 000 new terms to SentiFul.

During the course of this thesis the author explored ways to create sentiment lexicons from scratch and expand them automatically to improve their coverage. The following section describes these efforts.

## Lexicon Creation with Bootstrapping

One option to create sentiment lexicons is by expert decision. While being highly accurate this strategy is also time- and cost-intensive. Another option is to employ crowd-sourcing, e.g. via existing platforms such as Amazon's Mechanical Turk<sup>1</sup> or CrowdFlower<sup>2</sup>. However, the lack of intrinsic incentive results in lower motivation to complete the task with the required high quality.

This problem calls for an unprecedented solution. *Games with a purpose* offer such a solution. Designing the task as a game guarantees high motivation levels and also creates side-effects such as word-of-mouth to attract new players.

Invented by Luis von Ahn, games with a purpose have served for a variety of tasks, e.g. image recognition (von Ahn, 2006), annotation tasks (Siorpaes and Hepp, 2008), teaching robots (Kunze et al., 2013), or assessing the climate change awareness of the general public (Seebauer, 2013). The main idea is to design a task in a way to pretend that the person working on the task is actually playing a game. The work described in the following sections applies this principle to create an initial sentiment lexicon using a Facebook game, the so-called *Sentiment Quiz* (Scharl et al., 2012). The Sentiment Quiz was available in seven different languages (English, German, French, Spanish, Italian, Portuguese, and Russian) and attracted the interest of 4 300 players, who collected 1 000 high-quality terms for English as a side-effect of playing the game.

After the creation of the initial seed lexicon the subsequent application of a bootstrapping approach added further sentiment terms (Weichselbraun et al., 2011). This bootstrapping approach identified highly positive and negative reviews in a review corpus. Analyzing the occurrence frequencies of the terms in these strongly positive and negative reviews helped to identify further sentiment terms, which were not available in the initial seed lexicon. Including newly identified sentiment terms into the seed lexicon improved the coverage of the lexicon. A formal evaluation for the English language showed that the approach is promising. Thus, it is a promising approach for languages where resources in sentiment analysis are sparse and need to be created without expert annotators available to create them. The following sections describe the setup of the game, the strategies to avoid cheating among the players, and the bootstrapping procedure. The subsequent evaluation shows the efficacy of the approach.

### The Setup of the Sentiment Quiz

The Sentiment Quiz (Rafelsberger and Scharl, 2009) is a so-called *Game-with-a-purpose* (von Ahn, 2006), i.e. a game designed in a value-adding way, that models a sentiment annotation task as a game. The system presents potential terms to the players and asks for their opinion on the polarity and strength of the sentiment term. Players score when their answers match the answers of previous players. In cases where no prior answer is available players get scores equal to their average game performance. The game was implemented as a Facebook application to get access

---

<sup>1</sup><https://www.mturk.com/mturk/welcome>, last accessed on 24 November 2014.

<sup>2</sup><http://www.crowdfLOWER.com/>, last accessed on 24 November 2014.

to Facebook's large user community (see Figure 2.1). The players rate each term on a five-point scale.

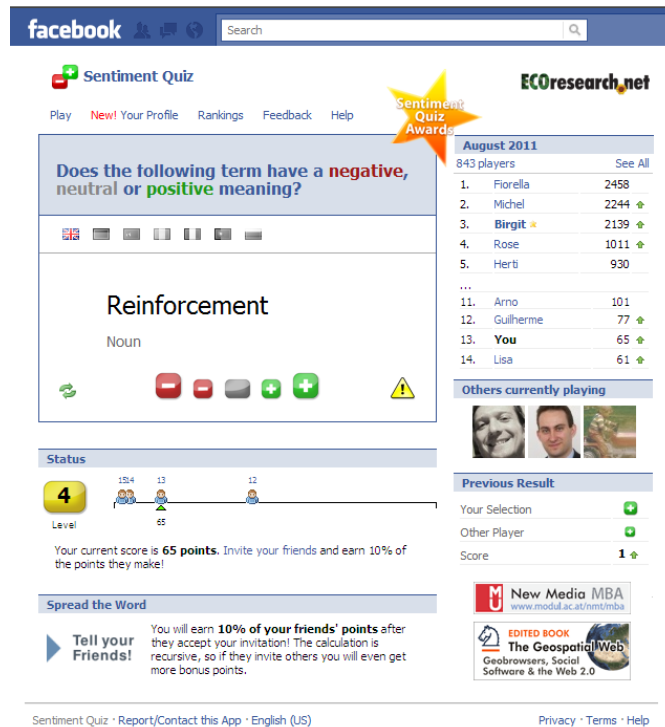


Figure 2.1: The Sentiment Quiz was deployed as a Facebook application to increase user engagement.

## Quality Management

Cheating is a widely known problem for all varieties of games. For instance, on a platform such as Facebook it is easy for players to communicate with each other and synchronize their answers, resulting in more points for the cheaters. Furthermore, randomized clicking or repeatedly clicking the same answer renders the data useless. The Sentiment Quiz employs a battery of strategies to combat cheating:

- **Cloaking:** make players invisible to each other to avoid player-to-player communication.
- **Answer analysis:** extract answer patterns, e.g. repeated clicking on one and the same answer option.
- **Credibility check:** assign credibility scores for users with a high number of correct responses.
- **Randomization:** avoid patterns in the game.

In addition to these strategies, each term needs a minimum of seven similar independent assessments before its inclusion into the sentiment lexicon.

### The Bootstrapping Procedure

After creating an initial sentiment lexicon containing 500 positive and 500 negative terms a bootstrapping method helped to expand this lexicon. The bootstrapping procedure consisted of three steps: (i) polarity annotation of unlabeled reviews; (ii) compilation of a sub-corpus consisting of the reviews with the strongest polarity; (iii) the extraction of unknown sentiment terms from this sub-corpus (also see Figure 2.2). The following sections describe each step in more detail.

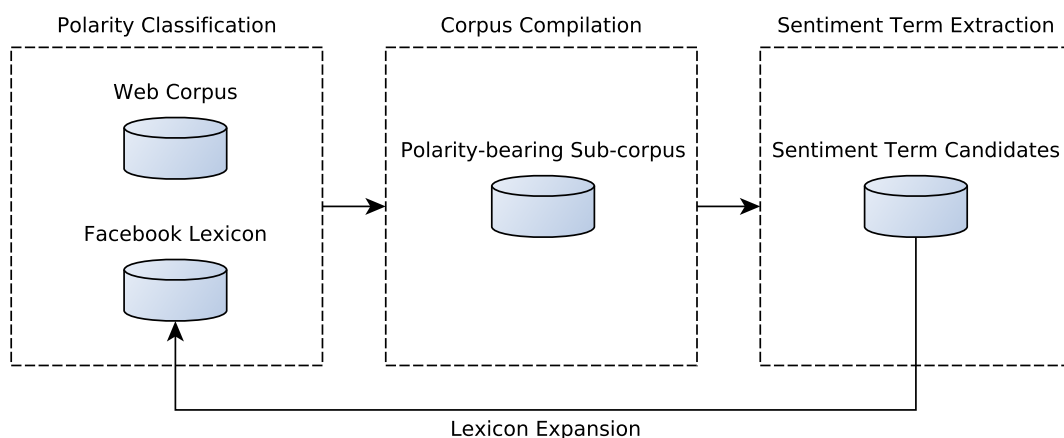


Figure 2.2: Overview of the bootstrapping procedure.

### Polarity Classification

The initial step of the bootstrapping procedure is the polarity classification of the reviews. Each review gets a polarity value assigned by using a keyword lookup algorithm with negation detection that uses the following formula:

$$\sigma(doc_i) = \sum_{t_j \in doc_i} n(t_{j-1})\sigma(t_j), \text{ with} \quad (2.1)$$

$$n(t_{i-j}) = \begin{cases} -1.0 & \text{if } t_{j-1} \text{ is a negation trigger} \\ +1.0 & \text{otherwise} \end{cases} \quad (2.2)$$

The used algorithm assigns a numeric label to each review reflecting its polarity and strength. This allows to apply a ranking, which helps to compile a sub-corpus of strong polar reviews in the next step.

## Compiling the Corpus

After assigning sentiment values to the reviews, the system extracts the reviews with the highest polarity values and uses them as a learning corpus for the extraction of new sentiment terms. The strength thresholds  $\sigma_k^+$  and  $\sigma_k^-$  serve as criteria for the inclusion of positive ( $C^+$ ) and negative ( $C^-$ ) reviews into this learning corpus:

$$C^+ = \{doc_i | \sigma(doc_i) > \sigma_k^+\} \quad (2.3)$$

$$C^- = \{doc_i | \sigma(doc_i) < \sigma_k^-\} \quad (2.4)$$

## Extracting New Sentiment Terms

Using the Naïve Bayes formula, the systems finally extract new sentiment terms based on their probability to occur in positive/negative reviews:

$$n(t_j) = n(t_j|C^+) + n(t_j|C^-) \quad (2.5)$$

$$P(\sigma(t_j)|C^+) = \frac{n(t_j|C^+)}{n(t_j)} \quad (2.6)$$

$$P(\sigma(t_j)|C^-) = \frac{n(t_j|C^-)}{n(t_j)} \quad (2.7)$$

Ranking the terms by probability strength and applying another integration threshold results in the inclusion of the strongest positive/negative sentiment terms:

$$\sigma(t_j) := 1 \quad \text{if } P(\sigma(t_j)|C^+) > P^+ \wedge n(t_j) \geq n_{min} \quad (2.8)$$

$$\sigma(t_j) := -1 \quad \text{if } P(\sigma(t_j)|C^-) > P^- \wedge n(t_j) \geq n_{min} \quad (2.9)$$

## Evaluation

The evaluation answered two questions:

- Does the bootstrapping procedure improve the overall quality of the initial lexicon?
- Is the efficacy of the created lexicon comparable to existing lexicons?

To answer the first question the evaluation compared the results of the Facebook lexicon with the expanded lexicon. Both lexicons served as input for (i) a keyword lookup algorithm, (ii) a Naïve Bayes classifier, and (iii) a Support Vector Machine, the latter two using the implementations of the WEKA toolkit (Hall et al., 2009).



For the second question, we compared the expanded lexicon with a sentiment lexicon derived from the sentiment terms in the General Inquirer. Again, the lexicons served as input for the keyword lookup, the Naïve Bayes and the Support Vector Machine classifier.

Figure 2.3 contains a schematic overview of the evaluation design. The three lexicons serve as the features for the WEKA classifiers, requiring the reviews to be turned into WEKA’s preferred ARFF format. They also serve as the input for the keyword algorithm.

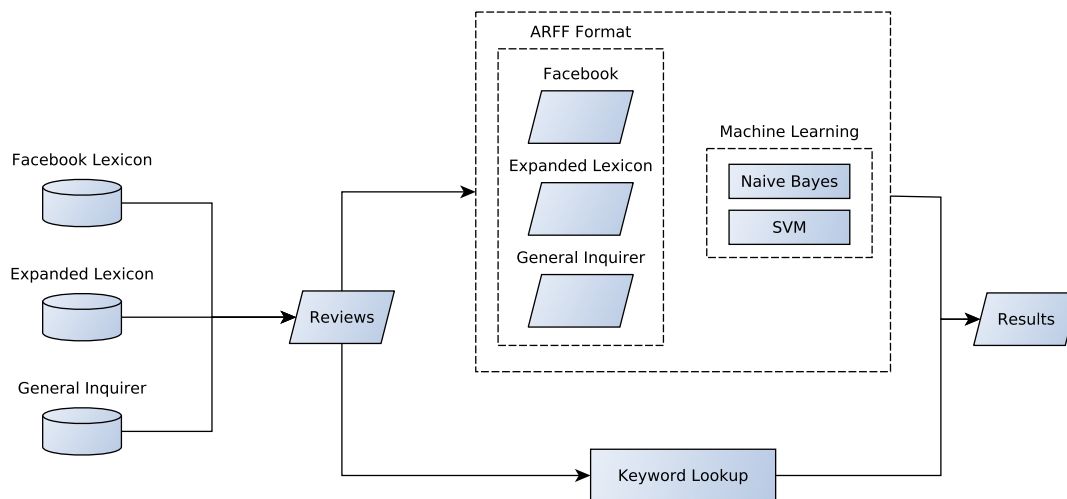


Figure 2.3: The evaluation compares the three lexicons using keyword lookup, Naïve Bayes and an SVM classifier.

The expanded lexicon performed considerably better than the initial lexicon. Table 2.1 contains the results of the 10-fold cross-validation (for an explanation please refer to Section 5.3), with 16 significant improvements of the expanded lexicon according to Wilcoxon’s rank sum test with  $p < 0.05$ . This indicates that the proposed strategy successfully expands an existing small lexicon containing high-quality terms.

<b>Polarity</b>	<b>R<sub>f</sub></b>	<b>R<sub>e</sub></b>	<b>Sig</b>	<b>P<sub>f</sub></b>	<b>P<sub>e</sub></b>	<b>Sig</b>	<b>F<sub>f</sub></b>	<b>F<sub>e</sub></b>	<b>Sig</b>
<b>Keyword Lookup</b>									
Positive	77	90	↑	62	69	↑	68	78	↑
Negative	29	43	↑	85	92	↑	43	58	↑
<b>Naïve Bayes</b>									
Positive	63	76	↑	75	79	↑	68	77	↑
Negative	79	79	·	68	76	↑	73	78	↑
<b>SVM</b>									
Positive	73	80	↑	75	79	↑	74	79	↑
Negative	75	78	·	74	80	↑	74	79	↑

Table 2.1: Comparison of the initial lexicon with the expanded lexicon.

The 10-fold cross-validation for the comparison of the expanded lexicon with the General Inquirer lexicon was more diverse (see Table 2.2). The expanded lexicon performed significantly better for precision of positive reviews and recall of negative reviews, when using the keyword lookup algorithm. On the other hand, the lexicon derived from the General Inquirer performed significantly better in five results, i.e. recall, precision, and f-measure for positive reviews and precision and f-measure of negative reviews, all of them using the WEKA Naïve Bayes classifier. The remaining results were statistically non-significant changes. Given that the General Inquirer is almost twice the size of the expanded lexicon and crafted by experts, the results are still promising.

<b>Polarity</b>	<b>R<sub>e</sub></b>	<b>R<sub>gi</sub></b>	<b>Sig</b>	<b>P<sub>e</sub></b>	<b>P<sub>gi</sub></b>	<b>Sig</b>	<b>F<sub>e</sub></b>	<b>F<sub>gi</sub></b>	<b>Sig</b>
<b>Keyword Lookup</b>									
Positive	90	95	.	69	65	↑	78	77	.
Negative	43	36	↑	92	93	.	58	52	.
<b>Naïve Bayes</b>									
Positive	76	85	↓	79	82	↓	77	83	↓
Negative	79	81	.	76	85	↓	78	82	↓
<b>SVM</b>									
Positive	80	86	.	79	82	.	79	84	.
Negative	78	81	.	80	85	.	79	83	.

Table 2.2: Comparison of the expanded lexicon with the General Inquirer.

Table 2.3 contains terms that were added to the sentiment lexicon during the bootstrapping procedure. Interestingly, the term *stops*, was included as a positive term, although it is intuitively a negative term. However, in the tourism domain, *stops* refers to locations with access to public transportation. Living close to a *bus stop* is desirable, as it equals easy connection to important places of the destination. Further examples are *dingy* and *stained*. Both terms have a negative sentiment, as they refer to undesired characteristics of a hotel.

The examples show that the presented algorithm is capable of extracting meaningful sentiment terms, although the learning corpus was rather small with only 1 600 reviews. Applying the bootstrapping procedure to a significantly larger corpus to further expand the lexicon is left for future work.

## 2.3 Polarity Classification

Polarity classification aims at attaching polarity labels to text entities. Common approaches classify documents, sentence, phrases or terms into positive or negative. The weight or strength of polarity is a further decisive indicator. Certain terms, although expressing the same polarity, might have a stronger or less strong contribution. E.g. “good” is a common indicator to express positive sentiment; “excellent” has the same polarity but is significantly stronger than the former one.

Term	Sentence
<i>stops (pos)</i>	Also lovely that the tram <i>stops</i> were literally outside our front door as it was very snowy a day or two during our week.
	It's just about 5 minutes from Stephansplatz, the U-Bahn and various tram <i>stops</i> .
	The hotel is off a quiet street, but easily reached from the airport by the 'CAT' train and then a few <i>stops</i> on the U3 underground and then a short stroll from here.
<i>dingy (neg)</i>	The hotel itself was shabby, <i>dingy</i> and very dirty looking.
	The lobby is reached through a dark, <i>dingy</i> restaurant and one had to walk past the largest smelliest dog I had ever seen.
	Sadly, it was in the rafters, dark and <i>dingy</i> seeming.
<i>stained (neg)</i>	The walls of the room were also very scuffed and <i>stained</i> .
	Our "Executive Room" featured dirty, <i>stained</i> old chairs and a coffee tablet that would have looked more at home in a rubbish skip.
	<i>Stained</i> bedspreads, soiled carpeting, broken telephone, and terribly noisy.

Table 2.3: Examples of terms added after bootstrapping.

In their early work Hatzivassiloglou and McKeown (1997) propagate sentiment values of known adjectives onto unknown adjectives by exploiting relations between them. Relations can either arise from conjunction terms (e.g., *and*, *or*, *but*) or from morphological differences (*thoughtful* vs. *thoughtless*). The authors assume that conjunctions like *and* propagate the same sentiment value, whereas *but* and the morphological relationships transfer the inverse value. From these assumptions, they train a log-linear regression model to identify groups of adjectives with the same orientation. A clustering algorithm further attaches positive/negative labels to the groups.

Bollegala et al. (2013) present a thesaurus-based approach to overcome domain-specificity and apply it successfully to reviews of different product domains. Jiang et al. (2011) work with self-annotated tweets using a target-dependent approach. Given the shortness of tweets, they, for instance, classify tweets as positive/negative when they express positive or negative sentiments towards certain target words, such as the *iPhone* in "*I love the iPhone*".

Document-level sentiment analysis assigns an overall polarity to an entire document by aggregating the polarity values of phrases and sentences. A very popular application area are customer reviews, such as movie, product, or destination reviews. Using reviews has several advantages:

- Crawling them from the Web is cheap and allows the creation of corpora big enough for machine learning.
- Corpus annotation is not necessary, since the authors of reviews usually provide a summary of their opinion in the form of a rating, e.g. star or circle rating in Amazon and TripAdvisor reviews. A common standardization technique is to ascribe reviews with less

than three stars (circles, respectively) to the negative category and those above three stars to the positive.

- The subjective nature of the reviews is guaranteed. Other documents, e.g. news articles usually have to undergo a subjectivity detection to separate them from those with mere objective, i.e. factual, content.

Implementing a system for polarity classification requires a dataset to test and/or train the system. Review corpora provide a shortcut when self-annotated resources are not available. Using well-established corpora (Ding et al., 2008; Hu and Liu, 2004; Pang and Lee, 2004) allows for a comparison with existing approaches. The early work by Pang et al. (2002) applies three different machine learning approaches (Naïve Bayes, Support Vector Machines and Maximum Entropy Modeling) on movie reviews. The machine learning techniques are well-known from topic categorization, yet could not deliver as good results for sentiment analysis as they can for the categorization of topics. The authors conclude that subtle, for the classifiers inaccessible features cause these poor results. Turney (2002) also performs binary classification on product reviews. Similar to Hatzivassiloglou and McKeown (1997) a set of known sentiment terms builds the basis. The application of point-wise mutual information (PMI) and latent semantic analysis (LSA) extends this basis. Beineke et al. (2004) refine this approach by using a Naïve Bayes model. Using this model, they also learn new words on both a labeled corpus (1 400 movie reviews by Pang et al. (2002)) and an unlabeled corpus consisting of 27 886 reviews (Pang and Lee, 2005). They use both a small seed list with five positive and negative sentiment terms, as well as a larger list, where the terms *good*, *best*, *bad*, *boring*, and *dreadful* as well as their WordNet synonyms represent the seed list. The authors conclude that their presented approach outperforms previous approaches regarding classification accuracy and speed of computation (thanks to a modified processing method). More fine-grained approaches determine the exact number of stars provided by the review author (Pang and Lee, 2005). Three methods (one-vs-all, regression, and metric-labeling) based on Support Vector Machines accomplish this task. In order to prove that such a fine-grained analysis is meaningful the authors conducted a test with humans. This test showed that humans are indeed capable of determining an exact star rating, and not only of making a binary decision.

Dave et al. (2003) compare the efficacy of a simple term counting algorithm with different machine learning algorithms on book, movie, and music reviews. Interestingly, the simple algorithm shows a performance similar to the more sophisticated classifiers. Zhang et al. (2008) examines the relation between the expected sentiment of a review and its helpfulness to identify features suitable for estimating the helpfulness of a review. Three different types of features serve for classification using SVMs:

- **Lexical similarity:** how similar is a customer review to an editorial review or the technical description of the product?
- **Shallow syntactic features:** can the number of proper nouns, modal verbs, etc. serve as a predictor for helpfulness?
- **Lexical subjectivity:** can sentiment terms serve as predictors for helpfulness?

The test dataset consists of a number of reviews from the domains electronics (Canon and Sony products), books (topic: engineering) and movies. The evaluation results of the study shows only minor correlation between lexical similarity and helpfulness and lexical subjectivity and helpfulness. Subrahmanian and Reforgiato (2008) examine the impact of adverb-verb-adjective combinations. They define a number of axioms describing how they influence each-other. For example, the combination of an intensifying adverb preceding a positive verb has stronger sentiment than the positive verb on its own. Such combinations can have an impact on adjectives or adverb-adjective combinations. The sentence sentiment using these combinations differs from that using each particle isolated from each-other. In their experiments, Subrahmanian and Reforgiato (2008) achieved promising results on 200 news pages.

Reviews as training data have proven to be beneficial in other related areas as well. Wollmer et al. (2013) show that they can support training polarity classifiers for speech recognition. Trilla and Alias (2013) evaluate their classifier for a text-to-speech application on Tweets.

Nicholls and Song (2009) examine the impact of different part-of-speech tags by employing a maximum entropy classifier. They considered only adverbs, adjectives, verbs and nouns as relevant for sentiment analysis and assigned these categories different weights. According to their results adjectives and adverbs are the strongest sentiment conveyors, while verbs and nouns contribute only little. Kim et al. (2006) show how to extract pros or cons from reviews (i.e. sentences expressing positive or negative sentiment). A maximum entropy model is used to accomplish this task. Yu and Hatzivassiloglou (2003) present a multi-layer sentiment analysis system. In a first step, documents from the Wall Street Journal are separated into opinionated and factual (a Naïve Bayes classifier accomplishes this task; it has been trained on the meta-data available for each Wall Street Journal article, which is split into *Editorial*, *Letter to editor*, *Business*, and *News*). The collections of opinionated and factual documents are further used to classify fresh sentences. For that purpose, the authors invoke three methods, a similarity measurement (if an unknown sentence is more similar to the sentences in the opinionated collection it is also considered to be opinionated), a simple and a multiple Naïve Bayes classifier, both trained on the sentences contained in the collection of opinionated and factual documents. They use a sentiment lexicon, where a set of seed terms (proposed in Hatzivassiloglou and McKeown (1997)) has been expanded by statistical methods to identify further sentiment terms. The average per word log-likelihood scores serve as measurement for the overall sentiment of a single sentence.

Agarwal et al. (2009) examine methods to automatically determine the polarity of subjective phrases in the Multi-Perspective Question Answering (MPQA) corpus, a manually annotated corpus containing subjective phrases (Wiebe et al., 2005). Their system extracts sentences from the corpus and determines several features for each subjective phrase. Amongst others, they assign the phrases the mean of the pleasantness score, activeness score and imagery score, which are dimensions in the Dictionary of Affect and Language (Whissell, 1989). Inspired by the activation-evaluation space representation proposed by Cowie et al. (2001) they also propose a so called *norm*, mathematically combining those three scores. The activation-evaluation scoring is also applied to process chunk features. Here, the system determines chunks in the processed sentence. A subjective phrase having an overlap with a chunk is expanded, so that it also includes

the chunk. The evaluation of the procedure is accomplished using different feature combinations, showing that all proposed features contribute to the improvement compared to a baseline.

Resources such as the MPQA corpus (Wiebe et al., 2005) or the well-known movie review corpus (Pang and Lee, 2004) are essential to evaluate a sentiment analysis system. However, resources in languages other than English are still sparse. Thus, the author of this thesis contributed to the creation of such resources and co-annotated two corpora for German. The subsequent section describes these efforts.

## **Annotation of Resources for Opinion Mining**

Resource sparseness is a prevalent problem in opinion mining. The reasons are manifold: annotating a corpus large enough for machine learning is time-consuming, setting up an annotation scheme is difficult because the definition of opinion is vague, differentiating between opinionated and factual language is difficult, and language subtleties cause disagreement among annotators and lowers inter-annotator agreement. The most extensive resources exist for the English language, with several sentiment lexicons (Esuli and Sebastiani, 2006; Hu and Liu, 2004; Mohammad and Turney, 2013; Stone, 1966; Wilson et al., 2005) and annotated corpora (Hu and Liu, 2004; Jindal and Liu, 2008; Wiebe et al., 2005). Resources in other languages are sparser, although efforts have already been taken, e.g. German sentiment lexicons (Clematide and Klenner, 2010; Remus et al., 2010; Waltinger, 2010), an annotated German review corpus (Klinger and Cimiano, 2014) or a lexicon in Czech (Veselovská et al., 2014).

In collaboration with the Interest Group on German Sentiment Analysis (IGGSA, see Section 3.2) efforts were taken to reduce resource sparseness for the German language, by creating publicly available, annotated corpora for evaluation purposes. The first corpus is annotated on three layers of different granularity, the second serves as a benchmark corpus for a shared task on opinion holder and target extraction.

## **Reference Corpus for German Sentiment Analysis**

The MLSA corpus, the **multi-layered** reference corpus for German **sentiment analysis** (Clematide et al., 2012), contains 270 sentences in total and is publicly available<sup>3</sup>. The three layers cover sentence-level, word- and phrase-level, and expression-level annotation (see Figure 2.4).

### **Sentence-level Annotation**

This is the most-coarse grained annotation, where an overall value is assigned to each sentence. A sentence is either positive, negative, or neutral. Three annotators performed this annotation task. For subjectivity classification, i.e. “does the sentence express sentiment or not”, they achieved an inter-annotator agreement of 0.721. For polarity classification, i.e. “is the sentence positive or negative”, the inter-annotator agreement was 0.765. Fleiss’ kappa with average pairwise agreement served as measurement for inter-annotator agreement (Fleiss, 1981).

---

<sup>3</sup><https://sites.google.com/site/iggsahome/downloads>, last accessed on 24 November 2014.

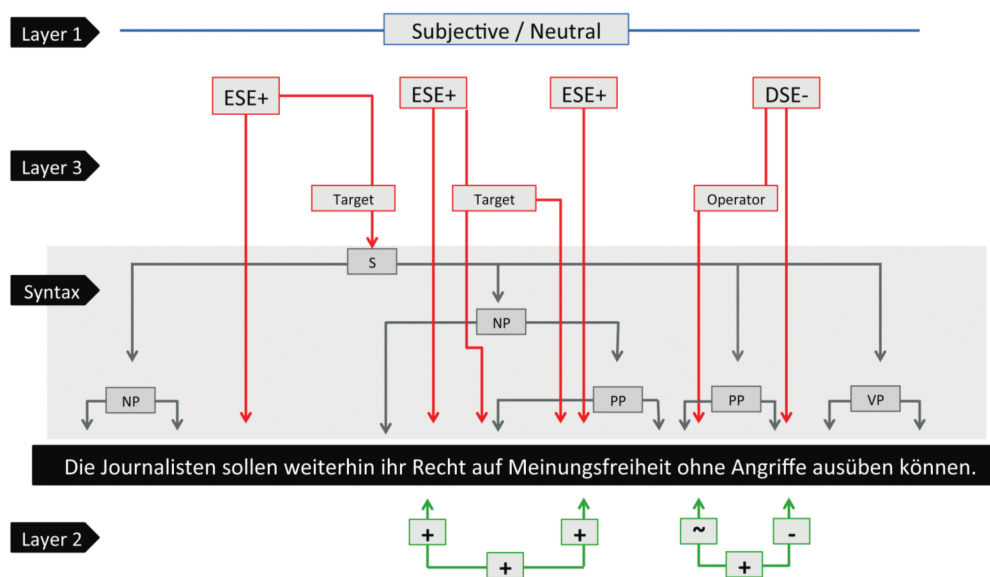


Figure 2.4: The three annotation layers on the example sentence “Journalists should continue to exercise their right to freedom of expression without attacks.”

### Word- and Phrase-level Annotation

The second layer is the annotation of words and phrases, i.e. noun phrases and prepositional phrases. In a first step, sentiment terms and modifiers receive a label. Sentiment terms are either positive (+), negative (−), neutral (0), or bipolar (#). Modifiers have three categories: shifters, inverting the sentiment value of a phrase (~), diminishers, lowering the strength (%), and intensifiers, increasing the strength (^). For example, for the sentence

*ohne Hass auf deine Peiniger*  
 [without hatred for your torturers]

the annotator first assigns polarity and modifier labels to single words:

*ohne~ Hass− auf deine Peiniger−*

Subsequently, entire phrases receive an overall sentiment value:

[*ohne~ Hass− [auf deine Peiniger−]−*]+

In the demonstrated example the polarity of the entire phrase shifts from negative to positive because of the shifter term *ohne*. The annotators achieved an agreement of 0.685 on the word level and 0.808 on the phrase level.

## Expression-level Annotation

The third layer is a resource for opinion holder and target extraction. The distinction between *objective speech events*, *direct speech events*, and *explicit speech events* allows the annotation of holders and targets, as shown in the following example (see Table 2.4 for a description of these labels):

[Peter]<sub>source</sub> [schimpfte]<sub>DSE</sub> [nicht]<sub>operator</sub> [viel]<sub>modulation</sub> [über das Wetter]<sub>target</sub>  
 [Peter]<sub>source</sub> does [not]<sub>operator</sub> [complain]<sub>DSE</sub> [much]<sub>modulation</sub> [about the weather]<sub>target</sub>

Name	Description and example
Direct speech event (DSE)	Speech particles directly address a target <i>Peter [schimpfte]<sub>DSE</sub>, über das Wetter.</i> <i>Peter [complained]<sub>DSE</sub> about the weather.</i>
Explicit speech event (ESE)	Statements of sentiment without explicit expression <i>Peter [sagte]<sub>OSE</sub>, dass es regnete.</i> <i>Peter [said]<sub>OSE</sub> it was raining.</i>
Objective speech event (OSE)	Statements expressing factual information <i>Peter [trägt]<sub>ESE</sub> eine furchtbare Jacke.</i> <i>Peter [wears]<sub>ESE</sub> a terrible jacket.</i>

Table 2.4: The three labels used for expression-level annotation.

## Shared Task on Source and Target Extraction from Political Speeches

Resource sparseness for tasks in opinion holder and target extraction motivated IGGSA to create a respective corpus (Ruppenhofer et al., 2014) and make it publicly available in a shared task collocated with KONVENS 2014, the Conference on Natural Language Processing (*Konferenz zur Verarbeitung Natürlicher Sprache*). Shared tasks are competitions where researchers in a field apply tools they developed on a standardized corpus. This helps to benchmark the tools and provides the community with a clear overview of the state of the art.

The created corpus consisted of 250 sentences extracted from speeches of the Swiss parliament. Swiss German is different from standard German not only in accent but also in the usage of certain terms, e.g. *vorprellen* instead of *vorpreschen* (*to press ahead*), the latter being used in Germany. To avoid these differences we focused on sentences dealing with non-parochial topics, i.e. topics not concentrating on locally limited affairs. The following reasons motivated the usage of political speeches:

- The data is publicly available.
- The texts are well-written and contain multiple sources and targets, making the annotation more interesting.
- Personal professional interest of the contributing researchers.



We used the following preprocessing pipeline to convert the raw data into its final format usable for the SALTO annotation tool (Burchardt et al., 2006):

1. OpenNLP for sentence segmentation and tokenization<sup>4</sup>.
2. The TreeTagger for lemmatization (Schmid, 1994).
3. The Berkeley parser for constituency parsing (Petrov and Klein, 2007).
4. The TIGER tools to convert the parse tree into the TIGER XML format (Lezius, 2002).

Creating a corpus for a shared task is a sensible procedure and requires a clear and accurate annotation scheme given the high complexity of human language. Figure 2.5 is the annotation of a sentence with a single opinion holder and target. However, the data structure is complex, involving several noun phrases and prepositional phrases, as can be seen in the parse tree. Figure 2.6 is a more complex example with three holders and targets in total. The complexity of such an annotation task requires the assistance of trained persons.

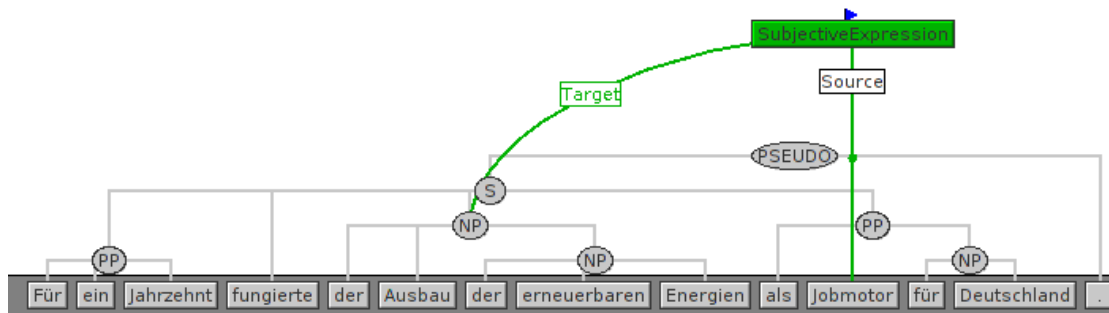


Figure 2.5: Example sentence with a single opinion holder and target.

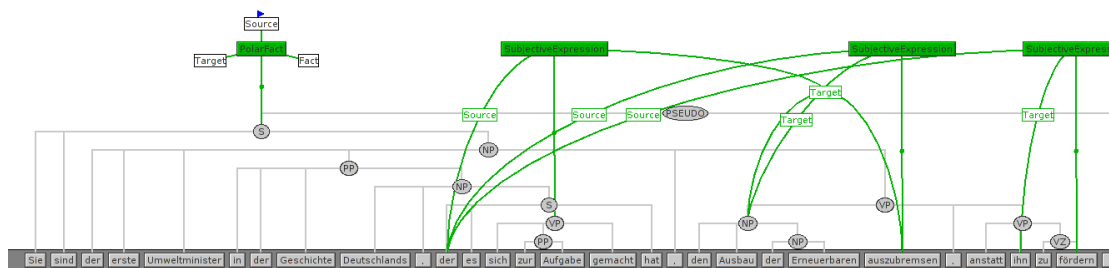


Figure 2.6: Multiple holders and targets make the annotation complex.

The creation of the annotation scheme started with an initial annotation of 50 sentences, accomplished by four annotators, to get an overview of the data characteristics. This step resulted in the creation of a set of initial guidelines. Re-annotating the sentences using these guidelines,

<sup>4</sup><http://opennlp.apache.org/>, last accessed on 24 November 2014.

accomplished by two annotators, validated and consolidated the annotation scheme. Subsequently, the remaining sentences were split into halves. Two teams of three annotators, one experienced annotator and two master-level students, annotated each half of the data. The first team annotated between 145 to 262 expressions as subjective, while the second team annotated between 122 and 236 expressions. Calculating the pairwise inter-annotator agreement showed that the average agreement was 0.62 for opinion sources, 0.65 for opinion targets, and 0.57 for subjective expressions.

## 2.4 Opinion Target Extraction

This sub-area of opinion mining answers the question “Who thinks what about whom/what”. Polarity classification and emotion analysis give valuable insights in the general sentiment and, respectively, the feeling of text instances but they do not deliver information about the interaction between them. Collecting this information is valuable for companies and enterprises but is usually too costly to achieve manually, which calls for automated approaches (Morinaga et al., 2002). For instance, the launch of a new product is a highly sensible phase for a company. Product flaws such as undetected software bugs or weaknesses in the used materials bear the potential for significant negative response and call for early-warning systems monitoring media response. Classical user tests or questionnaires are cost- and time-intensive, and biased by the beliefs of the conducting researchers. It is hard to prepare for all eventualities. For example, the detection of the iPhone 4 antenna problems were customer-driven and discovered after the product was shipped. Launching an iPhone creates massive media response, which facilitates the collection of feedback. In cases where the product is launched more silently it might take longer until momentum is sufficient to transport the negative opinions to the responsible persons. In the worst case it remains silently undetected and constantly subtracts new customers due to existing, but invisible negative word-of-mouth.

The technology is also highly beneficial for the planning and supervision of political campaigns and for public opinion monitoring during events of public interest, e.g. upcoming strikes because of unacceptable labor conditions, revolts caused by poor political decisions, or even climate change. Such systems help to detect problems before they get out of control. Detecting negative word-of-mouth is also highly beneficial in the political area.

An opinion mining toolkit capable of opinion target identification and differentiation of aspects, as presented in the following sections, is a powerful tool to handle exactly these problems. The presented unsupervised approach is capable of revealing hidden statements that would have been left undiscovered without this technology.

Summarizingly, opinion holder and target extraction involves the identification of three entities:

- **The opinion holder:** An entity expressing the opinion, e.g. a person, movie character, political agent, company, etc.
- **The opinion target:** The entity affected by the opinion.

- **The opinion aspect:** A feature of the target that further narrows down the expressed opinion, e.g. the lens of a camera or the haircut of a person.

In ideal cases all of these three are explicitly given in a sentence:

[The reviewer]<sub>holder</sub> thinks that [the lens]<sub>aspect</sub> of [the camera]<sub>target</sub> is inferior.

The same opinion can be expressed in a more subtle way, e.g.:

[The lens]<sub>aspect</sub> of [the camera]<sub>target</sub> is inferior.

The explicit opinion holder “reviewer” is not apparent in this sentence and has to be derived from the context. It is possible to leave out even more text particles without losing the actual information:

[The lens]<sub>aspect</sub> sucks.

The opinion target is not apparent in this sentence anymore and also needs to be derived from the context. Removing further information from the sentence makes the correct interpretation even more difficult:

It only makes blurry pictures.

Deriving the information of an inferior camera lens is only possible via the detour of “blurry pictures”, which requires a deep understanding of language and connections of the involved concepts. The examples are fictional but accurately reflect the circumstances in a current domain for sentiment analysis, i.e. customer reviews. Xueke et al. (2013) work with customer reviews and jointly extract opinion aspects and their sentiment given the respective domain. For restaurant reviews, this yields factual aspect, e.g. “table”, “reservation”, or “waiter”, negative aspects such as “rude”, “cold”, or “unfriendly”, and positive aspect, e.g. “friendly”, “great”, “nice”. Garcia-Moya et al. (2013) apply an approach based on language models to extract aspects such as “picture”, “resolution”, or “shot” in camera reviews, and Hai et al. (2014) identify “screen” and “battery” in a review sample on iPhone 5. Comparing aspect statistics in corpora of different domains creates indicators for their aspect extraction. Jakob and Gurevych (2010a) use conditional random fields to extract opinion targets from annotated review corpora of different domains to investigate their cross-domain applicability. Furthermore, they demonstrate the relevance of anaphora resolution (a sub-category of Co-reference resolution) for opinion target extraction in movie reviews (Jakob and Gurevych, 2010b). A potential application of such technologies is an automatic review summarization, as shown by Wang et al. (2013b). Their tool SumView extracts relevant features from a review corpus and provides a visualization of the summary for the interested user. SentiView follows a similar approach and visualizes temporal changes of the public opinion on common topics (Wang et al., 2013a). A significant amount of research is available for English, but there are also contributions in other languages: Klinger and Cimiano (2014) created a corpus for opinion aspect extraction in product reviews for German,

Zhu et al. (2011) demonstrate an unsupervised approach for Chinese restaurant reviews, and Wang et al. (2013c) mine implicit aspects in Chinese product reviews.

Micro-blogging services such as Twitter provide another valuable resource for opinion mining. The limitation in this domain are privacy issues, causing active Twitter users to keep their tweets non-public. Ren and Wu (2013) overcome this obstacle by applying the *homophily* theory, indicating that users with similar characteristics share similar opinions (McPherson et al., 2001). In contrast to these privacy limitations, the hashtags, a peculiarity in micro-blogs with the very distinct hash notation, offer additional information. For instance, Wang et al. (2011) exploit hashtags to assess the sentiment towards topics such as iPhone or Lady Gaga.

Opinion aspect extraction requires strong linguistic preprocessing. Techniques range from semantic role labeling combined with anaphora resolution (Ruppenhofer et al., 2008), conditional random fields (Jakob and Gurevych, 2010a; Nakagawa et al., 2010), dependency parsing (Nakagawa et al., 2010), or syntactical relations (Qiu et al., 2011; Sayeed et al., 2012). The double-propagation approach by Qiu et al. (2011) initially identifies opinion targets by connecting specific linguistic units (e.g. nouns) with sentiment terms using syntactical rules. In a second round, the rules connect further terms to targets of the first round. These new terms are either new targets or so far unknown sentiment terms, usable for the expansion of the used sentiment lexicon. The opinion aspect extraction approach as outlined later 2.4 relies on the relations defined by Qiu et al. (2011), but does not implement double-propagation. Instead, anaphora resolution as used by Jakob and Gurevych (2010b) helps to further improve target extraction. To overcome the obstacle of sparse resources for anaphora resolution (Charniak and Elsnar, 2009), they extend the existing tools MARS (Mitkov, 1998) and CogNIAC (Baldwin, 1997).

Yi et al. (2003) accomplish opinion holder and target detection on reviews of digital cameras and music. First, a part-of-speech tagger processes the sentence and extracts all phrases with at least one noun. Afterwards, they apply the mixture language model by Zhai and Lafferty (2001) and the likelihood-ratio test by Dunning (1993) to refine the selection and returns features that are targets. Sentiment analysis is performed on these extracted features. Two data sources serve as basis. On the one hand the authors use a sentiment lexicon, which is compiled from the sentiment terms in the General Inquirer (Stone, 1966) and the Dictionary of Affect and Language, (Whissell, 1989), which has been expanded with synonym terms available in WordNet (Fellbaum, 1998). On the other hand, they use a sentiment pattern database, containing verb patterns. Verbs can either directly affect a target (e.g. ‘impress’) or transfer sentiment from another object to the target (e.g. ‘be’). A verb’s property is also retrieved from the General Inquirer, the Dictionary of Affect and Language, and the emotion cluster of WordNet. All sentences containing a target are extracted from the collection. Among these sentences only kernel sentences, containing only one verb, are used. Subsequently, a parser identifies phrase chunks. Each chunk is assigned the sentiment value of the sentiment terms occurring in it. Via the verb sentiment patterns the chunks’ sentiment value is assigned to the target in the sentence. When sentiment patterns are missing (e.g. due to incompleteness of the sentences), the sentiment values of the chunks are directly assigned to the targets.

Hu and Liu (2004) extract features from product reviews (crawled from Amazon and CNET). In the second step they extract all sentences from reviews containing relevant features, and examine their subjectivity. A simple algorithm processes all subjective sentences and afterwards

the system creates for each product a list containing all found features as well as the number of positive and negative occurrences. Again, a sentiment lexicon is the basis for the algorithm. This is originally a seed list. For all terms in the seed list synonyms and antonyms are looked up in WordNet. Synonyms are integrated into the lexicon with the same sentiment value as the seed term, antonyms obtain the inverted sentiment value.

Popescu and Etzioni (2005) present Opine, a system for autonomous identification of product features and the assessment of the opinions available for those features. The whole process undergoes four steps, (1) product feature identification, (2) identification of opinions related to these features, (3) sentiment detection on the opinions, and (4) the ranking of the opinion strength. In the first step, Opine identifies relevant features by filtering nouns having a frequency higher than an empirically determined threshold. Afterwards, a rule base decides if there are opinions related to the features. In the next step, the system determines the sentimental orientation of the related opinions using relaxation labelling, a technique common in computer vision. Relaxation labelling is based on the assumption, that terms in the neighborhood of known strong sentiment terms also tend to express the same sentiment as the known term. The authors define *neighborhood* from a syntactical point of view. For example, *conjunctions* or *disjunctions*, *synonymy*, *antonymy* or *IS-A* relationships (the latter three derived from Wordnet<sup>5</sup>) can serve as neighborhood features. The last step comprises the ranking of the opinion strength.

Opinion Observer, developed by Liu et al. (2005), provides a visualization component, allowing a sentiment comparison of product features. The user of the system has, for example, the opportunity to compare the features of two cameras. The authors' example shows the superiority of one camera by contrasting the sentiment the cameras' features gained (see Figure 2.7). Opinion Observer can determine sentiment for features automatically or semi-automatically. Both the automatic as the semi-automatic version rely on sentences preclassified into pros and cons by the review author.

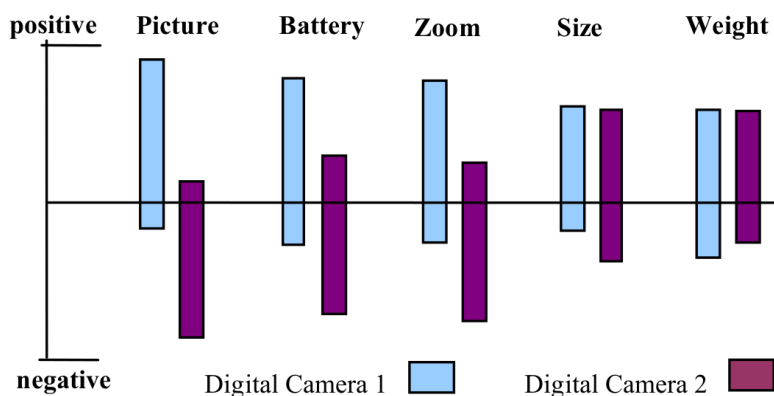


Figure 2.7: Opinion Observer neatly contrasts positive/negative sentiment towards aspects of digital cameras.

A more sophisticated option is not only to identify the target of an opinion but also the source (e.g. a person; this source is the so-called *opinion holder*). Kim and Hovy (2004) present

<sup>5</sup>see also: <http://wordnet.princeton.edu/>, last accessed on 24 November 2014.

such an approach. A named entity recognition accomplishes opinion holder identification - an opinion holder can either be a person or an organization. The sentiment detection component is a lexical approach. Again, a number of seed terms serves as basis. WordNet's synonymy and antonymy relations are the means to extend the basic lexicon. Starting with 44 sentiment verbs and 34 sentiment adjectives this procedure ends up with several thousands of sentiment verbs and adjectives. The authors propose three models for sentiment analysis: (i) a model where several negatives cancel each other out (with two negative terms in a sentence the system assigns a positive sentiment value to this sentence); the other two models assign sentences the (ii) harmonic or (iii) geometric mean of the sentiment terms' strengths in the related region. The authors experimented with different region sizes, ranging from the whole sentence to only a number of words between the opinion holder and the topic (i.e. the target of the opinion holder) of the sentence.

### **Rule-based Opinion Target Extraction**

We implemented an artifact applying grammar rules onto linguistically preprocessed sentences to identify the opinion targets (Gindl et al., 2013). Furthermore, a heuristic helped to differentiate between *targets* and *aspects*. A target is an entity receiving a multi-faceted statement by the author of a written statement. The different facets of the statements are the entity's aspects. For instance, a camera, the target, might have favorable and unfavorable aspects. In our context, aspects are components of a target, e.g. the lens of the camera or its battery. The principle is not tied to physical objects but can be applied to intangible entities as well, e.g. residing in a hotel might have the aspect *friendliness* or *cleanliness*.

A sub-set of the grammatical rules presented by Qiu et al. (2011) builds the background to transfer polarity charges from sentiment terms onto their targets. The Stanford parser (Rafferty and Manning, 2008) creates the dependency tree of the input, which is the basis for the grammar rules. The system applies the grammar rules to the parse tree of the sentence and subsequently extracts opinion targets.

### **Linguistic Rules**

Qiu et al. (2011) define their approach as *double-propagation*. The rationale behind this name is the idea of identifying unknown sentiment terms via the double propagation of sentiment values. In the first step, their system identifies opinion targets using a set of grammar rules. The grammar rules consider nouns connected with a sentiment term via specific grammatical structures as opinion targets. In a second step, they use additional grammar rules connecting identified targets with potential new sentiment terms. Their hypothesis assumes that a term connected to an already detected target carries a sentiment charge itself, and should be integrated into the underlying sentiment lexicon if it is missing.

We build upon this work by adopting two of the used rules, combining it with a heuristic to bridge sentence borders (Lau et al., 2009) and completing it with a set of regular expressions to differentiate between targets and aspects. We did not invoke the double propagation procedure, as this work primarily focused on the extraction of targets and aspects.

The first rule transfers a polarity charge from an opinionated term  $O$  onto a target  $T$ :

$$O \rightarrow O - Dep \rightarrow T,$$

$$\text{s.t. } O \in \{O\}, O - Dep \in \{MR\}, POS(T) \in \{N, NN, NNP\}$$

s.t.  $\{O\}$  is the set of known sentiment terms and  $\{MR\}$  the set of used dependency relations:

- **advmod:** adverbial modifier
- **amod:** adjectival modifier
- **rcmod:** relative clause modifier
- **nsubj:** nominal subject
- **dobj:** direct object
- **nn:** noun compound modifier.

For instance, the rule connects the sentiment term *good* with the target term *phone* in the sentence

*The phone has a good screen.*

and transmits its positive charge onto the target. Figure 2.8 is a graphical representation of the propagation procedure. A force-directed placement algorithm with random initialization provided the locations of the term nodes, thus they do not represent any hierarchical order.

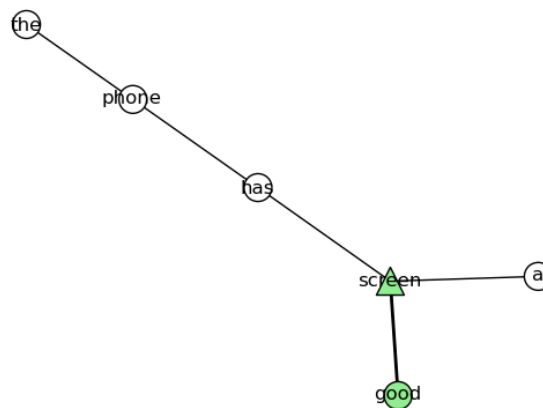


Figure 2.8: *The phone has a good screen* - *screen* receives a positive charge from *good*.

Modeling sentiment analysis as a mere polarity classification task suffers from reduced over-all charges, e.g. when a text snippet contains several sentiment terms with differing polarity. Target detection combats this problem by exactly pinpointing each target of a sentiment term. The sentence

*The phone has a good screen but a bad battery.*

contains the two sentiment terms *good* and *bad*, which have opposing polarities. Traditional polarity classification will assign a neutral overall sentiment value, since a positive and a negative value cancel each other out. Using the proposed approach overcomes this problem - the system identifies *screen* as a target with positive connotation and *battery* with negative connotation. This behavior allows fine-grained analysis of the entities mentioned in a text. Figure 2.9 demonstrates the detection of the two targets with differing polarity. Again, a force-directed placement algorithm positioned the term nodes, i.e. their location in the graph does not represent a hierarchical order.

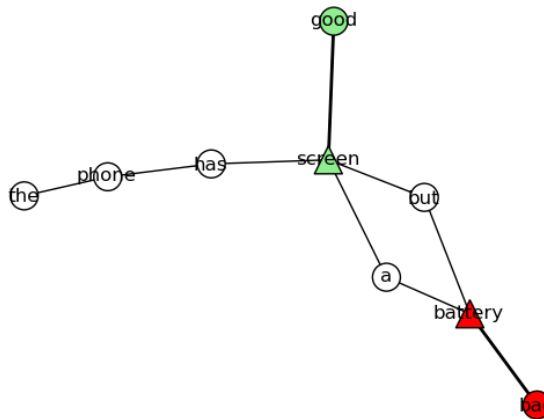


Figure 2.9: *The phone has a good screen but a bad battery* - multiple targets with differing polarities in one sentence.

The second rule transfers the sentiment charge of a target identified with the first rule onto another noun target.

$$O \rightarrow O - Dep \rightarrow H \leftarrow T - Dep \leftarrow T,$$

$$\text{s.t. } O \in \{O\}, O/T - Dep \in \{\text{MR}\}, \text{POS}(T) \in \{\text{N}\}, \text{NN, NNP}\}$$

For instance, the sentence

*The iPod is the best mp3 player.*

contains the target *player*, which is accessible via the first rule. Subsequently applying the second grammar rule also detects *iPod*, as this term is connected with *player* via the second grammar rule. The presented two grammar rules work within sentence boundaries, i.e. they do not transfer charges across sentences. To overcome these sentence boundaries a heuristic performs cross-sentence propagation and connects neighboring sentences.

For instance, the sentence



*Yesterday I bought a new phone. It is the best purchase I have ever made.*

contains the target *phone* in the first sentence. The second sentence has a reference to *phone* via the personal pronoun *it*. Merely using the discussed grammar rules cannot identify *phone* as a target, since they are limited to single sentences. To overcome this limitation, a heuristic bridges two sentences if they are connected via a personal pronoun (Lau et al., 2009). If the second sentence starts with a personal pronoun the approach assumes that this pronoun is a reference to the last noun in the previous sentence.

Figure 2.10 shows an example of this heuristic. The system identifies *purchase* as the target of the opinionated term *best* using the first grammar rule and *it* as a personal pronoun. Applying the second grammar rule transfers the positive charge onto *it*. Subsequently, the system connects *it* with the last noun of the previous sentence, i.e. *phone*, and transfers its positive value onto *phone*. Again, the locations of the term nodes do not represent any hierarchical order, because a force-directed placement algorithm was used to position them.

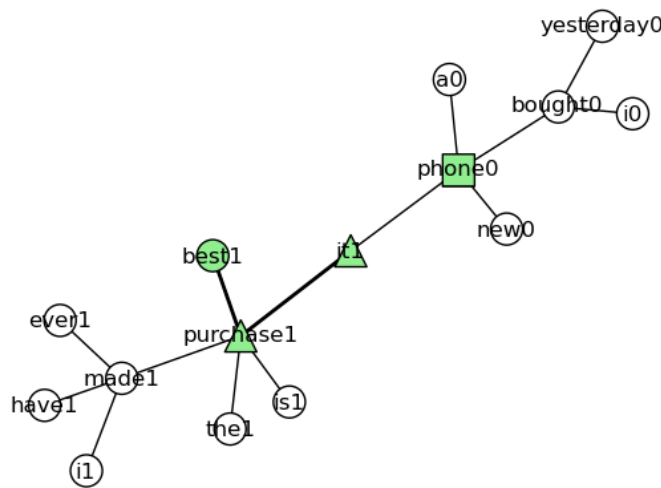


Figure 2.10: Identifying sentiment targets across two sentences.

Such a fine-grained approach allows for a highly detailed analysis of the opinion in a text. It replaces a mere binary classification with a contrasting summary of differing opinions towards one and the same target. This characteristic makes opinion target extraction an indispensable method in every sentiment analysis toolkit.

### Opinion Target vs. Opinion Aspect

The differentiation between opinion target and opinion aspect allows for a fine-grained analysis of the text corpus and facilitates the identification of hidden features. Such an approach reveals thus far unknown aspects of, for example, a product and reveals insights that are otherwise hard to get. In this work, an *opinion aspect* is a noun embedded in particular part-of-speech patterns (see Table 2.5), co-occurring with a previously identified opinion target. Implemented as regular expressions these patterns work with the Penn Treebank II tag set (Marcus et al., 1993).

Opinion Target Pattern	Part-of-Speech Pattern	Description
ADJECTIVE* NOUN+	(JJ (R S) ?) * (NN (S PS P) ?) +	adjectival noun phrase
ADVERB ADJECTIVE+ NOUN+	(RB (R S) ?) (JJ (R S) ?) + (NN (S PS P) ?) +	adverbial noun phrases
ADJECTIVE* NOUN PREPOSITION\	(JJ (R S) ?) * (NN (S PS P) ? IN) \	extended noun phrase
ADJECTIVE* NOUN+	(JJ (R S) ?) * (NN (S PS P) ?) +	

Table 2.5: Regular expression patterns for opinion aspect extraction.

## Analysis

The evaluation design compensates for the lack of appropriately annotated data. The implemented system extracted targets and their aspects from a large-scale Web corpus, consisting of 100 000 Amazon reviews. This corpus type is especially useful because of its guaranteed opinionated content. Furthermore, it contains descriptions of targets and aspects by nature, making it the logical choice for the task.

The data is not annotated, which aggravates the application of quantitative statistical measurements. As a workaround and proof-of-concept we analyzed the items extracted by the implemented system. Analyzing only items with high corpus frequency helped focus on the most relevant items and not to get lost in the sheer amount of data.

Table 2.6 contains the 15 most frequent positive and negative targets in the corpus. As one might expect people frequently talk about the *quality* and *price* of a product. *Sound*, *battery*, or *camera*, e.g. of a smart-phone, are further highly relevant targets. Interestingly, some targets are highly frequent as positive and negative targets alike. For instance, *quality*, *product*, and *sound* are frequent ambivalently discussed targets.

Positive targets	Negative targets
quality, 7534	quality, 2708
product, 6429	product, 2227
price, 4486	drive, 2043
sound, 4027	one, 1548
case, 3851	thing, 1505
one, 2350	battery, 1315
thing, 2302	case, 1219
camera, 2258	sound, 1115
picture, 1823	design, 1085
screen, 1805	time, 1072
value, 1624	screen, 1060
cable, 1549	cable, 929
battery, 1547	camera, 906
feature, 1388	unit, 905
device, 1330	software, 715

Table 2.6: The top 15 most frequent positive and negative targets with their respective frequency counts.

A similar evaluation design helped assessing the quality of aspect extraction. The system extracted meaningful aspects such as *sound quality*, *light weight*, or *low price* as positively discussed aspects. The extraction of negatively addressed aspects seems to be more problematic. The system extracted rather generic terms such as *first time*, *first one*, or *few days* as negative aspects. Further research will delve into these problems and investigate methods for obtaining more intuitive aspects. Table 2.7 lists the 15 aspects with the highest corpus frequency.

Positive	Negative
sound quality	first time
light weight	first one
high quality	other reviews
digital camera	few days
low price	second one
little camera	whole thing
small size	only problem
long battery life	few weeks
remote control	many times
build quality	few months
little device	second time
wide angle lens	big deal
extra money	only reason
audio quality	same thing
spare battery	few minutes

Table 2.7: Top 15 strongest positive and negative aspects.

Table 2.8 lists example sentences with both opinion targets and their aspect. In the first example the author praises a *webcam* for its high quality, *clear crisp photos*. The second example is a complaint about a weakly designed *power supply*, and the third example describes an intelligent container design, where the *metal box* itself contributes to *heat dissipation*.

Target	Aspect	Sentence
webcam (+)	crisp photos	i love the webcam work really well, clear crisp photos.
power supply (-)	wimpy feather-weight	Speaking of power, the Sabrent enclosure comes with a wimpy feather-weight 12V power supply rated at 2A bit I really doubt that is is capable of half that ...
box (+)	effective heat sink	The metal box itself is already a very effective heat sink for the drive.

Table 2.8: Example sentences with targets and aspects.

These examples show that the proposed approach is helpful for the detection of opinion targets. Furthermore, the approach extracts aspects, i.e. characteristics of the target, autonomously, allowing a further in-depth analysis. Without such a tool it becomes significantly harder to iden-

tify targets and their aspects without already knowing them. The approach is fully unsupervised, which omits the necessity of creating a training corpus. Rule adaption makes the approach readily applicable in other languages as well.

## Discussion

Opinion target extraction has significant business potential. Revealing negative aspects of a product is crucial for a company to improve the respective product. Discovering negative response for a newly launched product as quickly as possible does not only help to avoid similar future problems but also allows to develop media responses or compensation strategies for angry customers before the negative sentiment gets out of control. Particularly interesting is the detection of unexpected aspects, as it helps anticipating the unforeseen.

The detection of positive aspects is also highly interesting. Aligning the marketing strategy accordingly allows to emphasize the product aspects that are already well-received and get the best media response. Again, revealing what thus far was hidden, delivers insights in the product perception that are difficult to obtain otherwise. In cases where the gained evidence is strong enough a pivot in the product development strategy might become necessary.

The research area is also interesting from a scientific point of view, as it allows the exploitation of free-text corpora and integrate the gained knowledge into existing knowledge bases or create them from scratch. For instance, such an approach facilitates the creation of domain-specific, opinion-centered knowledge bases. Connected with topic identification tools they adapt to the underlying domain of an unknown document and consequently form a valuable chain of tools in an opinion mining toolkit.

Although opinion target extraction is a highly beneficial feature for sentiment analysis, it suffers from problems inherent in many Web documents. Even the best parsers have flaws, introducing mistakes in the detection of targets. Moreover, Web documents are often colloquial, which further reduces the accuracy of parsers.

## 2.5 Affect Analysis

The mere classification into positive/negative, or on a scale from  $[-1, 1]$ , respectively, misses a significant amount of the information conveyed in opinions. Opinions use a wide range of human emotions, which are openly neglected by these approaches. Ekman (1999) identifies at least six emotion inherent to humans. Analyzing the facial expressions of happiness, anger, fear, sadness, disgust, and surprise in photographs with people from 21 different countries showed a significant amount of overlap. Plutchik (2001) identifies eight different basic emotions, i.e. ecstasy, admiration, terror, amazement, grief, loathing, rage, and vigilance. Figure 2.11 depicts these emotional categories as petals of a blossom, also known as the *wheel of emotions*. Folding them results in a cone. The cone emphasizes the relevance of strength of emotions on the one hand as well as their relatedness. Going down the vertical axis of the cone reduces the strength of the emotion (e.g. from “rage” via “anger” to “annoyance”), while following the latitudinal circles leads from one connected emotion to the other (e.g. “ecstasy” is connected with “vigilance”).

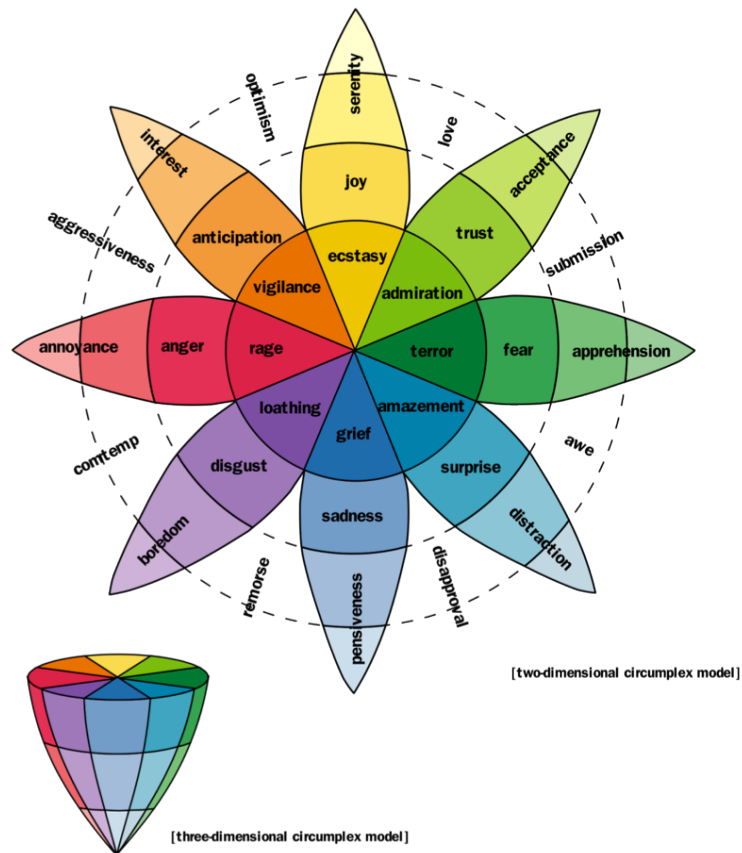


Figure 2.11: The wheel of emotions knows eight basic emotions.

Cambria et al. (2012) revise Plutchik’s wheel of emotions by interpreting emotions on opposed petals as belonging to one and the same abstract emotional class but positioned on the opposite side of the axis. For instance, “ecstasy” and “grief” are both emotions in the category “pleasantness” but on opposite sides of the spectrum. Their interpretation involves the four emotional axes “sensitivity”, “aptitude”, “attention”, and “pleasantness”, also called *sentic categories*. The resulting shape gives the model its name: *hourglass of emotions* (see Figure 2.12).

The *hourglass of emotions* is the theoretical background for SenticNet (Cambria et al., 2014). SenticNet contains polarity, sentics, and semantic knowledge for 50 000 concepts, thus combining common and common-sense knowledge in one resource. WordNet-Affect (Strapparava and Valitutti, 2004), another resource for emotion analysis, draws upon the knowledge contained in WordNet (Fellbaum, 1998). Assigning affective labels to WordNet synsets from the lexical

database AFFECT and subsequently leveraging synset relations to propagate affective labels onto further synsets provided affective data for 2 874 WordNet synsets and 4 787 words in total. Poria et al. (2013) merge SenticNet and WordNet-Affect to apply the affective labels of WordNet-Affect onto SenticNet concepts.

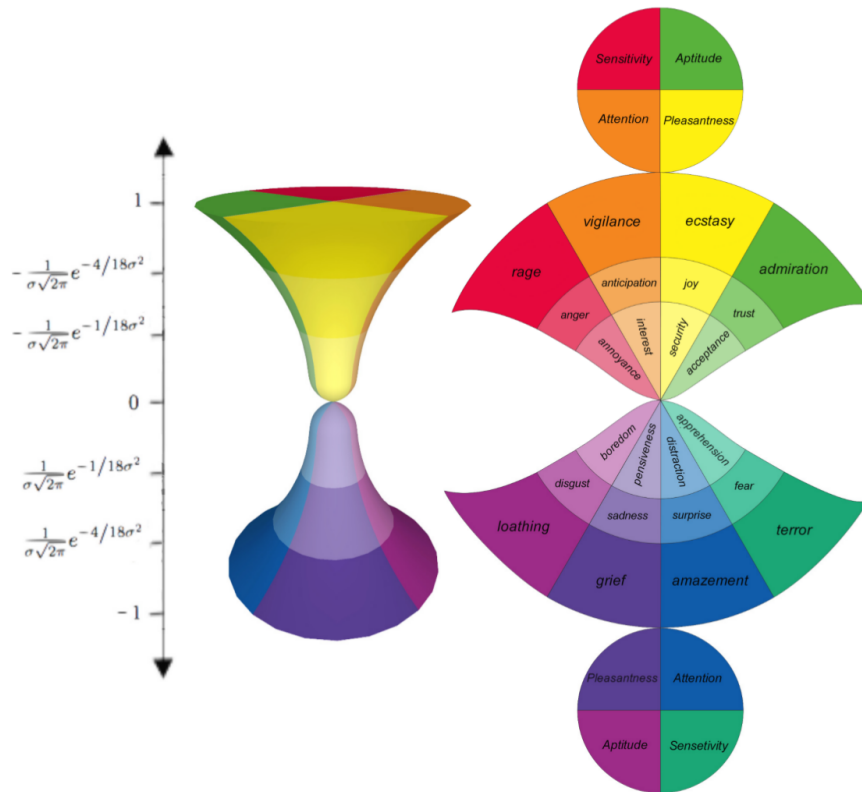


Figure 2.12: The hourglass of emotions.

Krcadinac et al. (2013) draw upon the aforementioned emotion classification by Ekman. Their open-source emotion recognition system “Synesketch” visualizes emotions, translating the emotional content of online communication into visual art. The visual sketch attempts to provoke similar emotions as conveyed by the communication thread (see Figure 2.13 for an example visualization of a Skype conversation).

The approach presented by Neviarouskaya and Aono (2013) refines emotion analysis by also taking context into account. They differentiate between unambiguous and equivocal terms, the first ones with fixed affective direction. The affective direction of the latter type of terms is adaptable depending on the context. Paltoglou et al. (2013) limit their approach to the two affective dimensions “valence” and “arousal” on a real-valued scale and evaluate it on forum discussions.

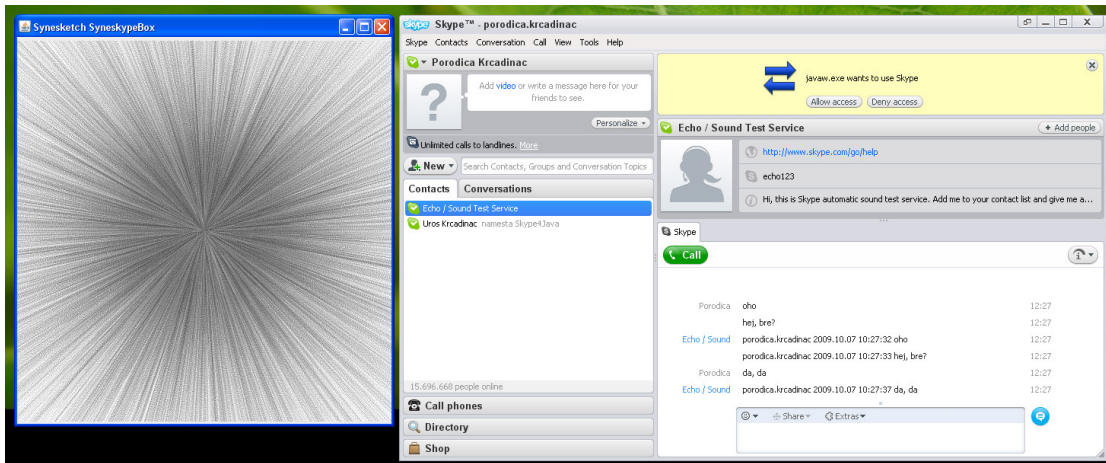


Figure 2.13: Visualization of a Skype conversation with Synesketch.

## 2.6 Invocation of Context Information

Context information is an essential ingredient for accurate sentiment analysis (Lau et al., 2009; Neviarouskaya and Aono, 2013; Wilson et al., 2005, 2009). One and the same term might have a different meaning in a different context or domain. For instance, the term *long* conveys positive sentiment in conjunction with *autonomy* (e.g. the autonomy of a laptop when not plugged in), while it expresses negative sentiment in conjunction with *delay* when travelling with an airplane (Cruz et al., 2013). Similarly, terms such as *large*, *small*, *high*, *low* are susceptible to polarity changes depending on their environment (Wu and Wen, 2010). Disambiguating them is a significant challenge in opinion mining and has already been attempted in competitive surroundings (e.g. as a Semeval tasks, such as by Lu and Tsou (2010), Xu et al. (2010), Yang and Liu (2010)). Domain adaptability, as it is also referred to (e.g. Remus (2012), Xia et al. (2013)), plays a significant role when applying classifiers to a domain different from their training domain. In the following we describe approaches leveraging contextual information to improve sentiment analysis. The presented paper also uses context information, made accessible in a new and innovative way.

Context helps improving approaches in all the three outlined sub-areas of sentiment analysis.

### Context in Polarity Classification

Lau et al. (2009) apply it to a polarity classification task and show its efficacy. Cross-domain sentiment classification also benefits from employing context as shown by Bollegala et al. (2013). Agarwal et al. (2009) accomplish phrase-level polarity detection refined by employing n-grams.

### Context in Emotion Analysis

Neviarouskaya and Aono (2013) classify sentiment words into the three attitude categories affect, judgment and appreciation. They analyze context using three different feature-selection al-

gorithms, i.e. point-wise mutual information, co-occurrence analysis, employed on Bing search results, and minimal path length employed on WordNet. Comparing Decision Tree and Naïve Bayes classifiers with Support Vector Machines showed that the latter delivered the best results for this task. Malandrakis et al. (2013) use context-based similarity metrics to generate affect labels for n-grams.

## Context in Opinion Holder and Target Extraction

Gangemi et al. (2014) underpin the relevance of context in opinion holder and target extraction. Context serves as an opinion trigger indicating that an entity expresses an opinion on a topic. Ren and Wu (2013) approach the problem in tweets with both social context (i.e. relations among users) and topical context (i.e. similar topics will share similar opinions).

According to Nasukawa and Yi (2003), sentiment detection consists of three steps: (1) the identification of sentiment expressions, (2) the determination of polarity and strength of the expressions and (3) the relationship of the sentiment expressions to their subject. Relationships are modelled in the treatment of verbs, which can either directly affect an argument (i.e. a target term) or transfer sentiment from one argument to the other. With such a model the authors are able to handle expressions like *XXX prevents trouble*. In that example, the verb *prevents* transfers the opposite sentiment of argument *trouble* to argument *XXX*. Terms with parts-of-speech different from ‘verb’ are treated in a simpler way - they directly transfer their sentiment to the related argument.

Wilson et al. (2005) present a two-step approach, first filtering polar sentences from neutral ones, and afterwards determining the polarity of the sentence. For the filtering process, context information is leveraged by invoking a total of 28 features, divided into five classes: (1) *word features* comprising regions around a sentiment expression, their part-of-speech but also the a priori polarity; (2) *modification features* include adjectives or adverbs changing the meaning of a sentiment expression, but also intensifiers (terms increasing the impact of an expression); (3) in *sentence features* sentences surrounding the currently analyzed sentence are also considered, but they also include the number of adjectives or adverbs in the current sentence; (4) *structure features* pay regard to the differentiation of active and passive voice; finally, (5) *document features* include a number of 15 possible document topics. The second step, the polarity classification, uses another two classes of totally ten features, (1) *word features* comprising a priori polarity, and (2) *polarity features* including negations. The features of both steps are trained and tested with BoosTexter’s (Schapire and Singer, 2000) AdaBoost.MH algorithm and identifies sentiment expressions in the MPQA corpus (Wiebe et al., 2005). The evaluation shows that both polar-neutral filtering and polarity classification benefits from using the proposed features. Wilson et al. (2009) expand this approach and use four different machine learning algorithms, BoosTexter’s Adaboost.HM, the rule-based learner Ripper (Cohen, 1996), TiMBL (Daelemans et al., 2001) for memory-based learning and the SVM implementations by Joachims (1999). For the evaluation of the system the authors use an extended part of the MPQA corpus. The findings of this work show that neutral-polar filtering is important and that large feature sets are necessary to accomplish both neutral-polar filtering as well as polarity classification.

Polanyi and Zaenen (2006) address several issues on context recognition and propose handling strategies from a linguistic point of view. They divide concepts responsible for context



switches into two groups: *sentence-based contextual valence shifters* and *discourse-based contextual valence shifters*. The first group can be divided into four subgroups:

- **Negatives and intensifiers:** negations invert the basic sentiment of a given sentiment term, as in “John **is** clever” vs. “John **is not** clever”. Intensifiers, such as *deeply* are capable of enhancing a term’s strength (e.g. “suspicious” vs. “deeply suspicious”). The authors also exemplify terms capable of diminishing a term’s strength, such as *rather* (e.g. “efficient” vs. “rather efficient”).
- **Modals:** shift sentiment by adding the concept of *possibility*, e.g. “Mary is a terrible person. She is mean to her dogs” vs. “If Mary were a terrible person, she would be mean to her dogs”.
- **Presuppositional items:** this group applies, when expectations are not met, e.g. “He *even* got into Harvard”.
- **Irony:** irony can completely turn around sentiment, e.g. “The very brilliant organizer failed to solve the problem”.

The second shifter group is divided into seven subgroups, e.g.:

- **Connectors:** terms connecting sub-sentences, such as *although*, *however*, or *but*. In the sentence “Although Boris is **brilliant** at math, he is a **horrible** teacher”, the term *although* connects the first (positive) subsentence with the second (negative) subsentence and neutralizes the positive sentiment of the first.
- **Discourse structure and attitude assessment:** the valence of one sentence spans following sentences, e.g. “John is a terrific athlete. Last week he walked 25 miles on Tuesdays. Wednesdays he walked another 25 miles.”
- **Genre constraints:** some genres are more difficult to classify correctly than others. For example, movie reviews often consist of the description of the plot as well as the author’s opinion on the film. These two parts have to be correctly distinguished to allow a proper sentiment detection.

The relevance of contextual valence shifters has successfully been shown in the literature, e.g. Neviarouskaya et al. (2011). SentiWordNet, a sentiment resource based on WordNet, was developed by Esuli and Sebastiani (2006). Their approach also uses context invocation by propagating sentiment values across synset terms. They use a semi-supervised approach to classify all WordNet synsets into positive, negative and objective. At first, they manually label all synsets containing 14 paradigmatic terms, creating 47 positive and 58 negative synsets. All synsets having a connection to these seed synsets are labeled accordingly. Used relations are *direct antonymy*, *similarity*, *derived-from*, *pertains-to*, *attribute*, and *also-see*. Afterwards, they identify objective synsets as those synsets which are not in the previously identified bag of synsets and which contain objective terms according to the General Inquirer. These three sets serve as

training data to train eight ternary classifiers, which then classify the remaining parts of WordNet. Lau et al. (2009) present an unsupervised approach for contextual sentiment detection. They use language models to rank opinionated web documents. Three different types of language models are used, a simple version based on a pre-defined query, a more complex version having a sentiment lexicon integrated and an inferential language model (Nie et al., 2006) capable of context involvement. According to the results the inferential language model clearly outperforms the simpler versions.

## Description of the Scientific Framework

*I think it's important to reason from first principles rather than by analogy. The normal way we conduct our lives is we reason by analogy. With analogy we are doing this because it's like something else that was done, or it is like what other people are doing. With first principles you boil things down to the most fundamental truths... and then reason up from there.*

Elon Musk.

The methodology used for this thesis relies on the principles of design science research (Hevner et al., 2004), a research area contributing to the field of information systems. Design science research strongly focuses on the design, creation, and evaluation of an artifact, useful enough to be implemented in a real-world environment, e.g. a business or company. Research in this area requires the adherence to certain standards, both from a design perspective (i.e. standards necessary to successfully create and evaluate an artifact) as well as from the knowledge transfer perspective, where the means of transferring gained insight to business stakeholders or the scientific community are relevant.

The following sections give an in-depth description of design science research based on the work of Hevner et al. (2004) as well as a connection between the principles of design science and the approach developed in this thesis. The presented approach suggests a solution for the problem of context-dependent polarity switches of sentiment-bearing terms. As described earlier, current systems in sentiment analysis rely on sentiment lexicons as their basis. The sentiment lexicon is a list of opinionated terms, i.e. terms that convey either positive or negative sentiment. The sentiment lexicon contains a-priori sentiment values for each term, e.g. the value -1 for the

term *terror*. While this a-priori value is a valid assumption in cases where no further information is available, it becomes problematic when the term is embedded in a context that changes the a-priori value. For instance, the term *repair* might be used when something is broken indicating negative context. On the other hand, after successfully repairing a physical object it becomes usable again, which indicates positive context. The approach of this thesis identifies sentiment terms susceptible for polarity shifts in a first step, considering them as “ambiguous terms”. Subsequently, it creates a contextualized sentiment lexicon, containing the ambiguous terms as well as a collection of co-occurring terms, the context terms. For each context term, the contextualized lexicon also stores the probability that this term, together with the ambiguous term, expresses positive/negative context.

In the application phase, i.e. when analyzing a free-text document, the system at first identifies ambiguous terms by checking if they have an entry in the contextualized lexicon. If this is the case, the system *disambiguates* the ambiguous term by considering the context terms of the ambiguous terms. It results in an assessment whether the ambiguous term is more likely to be positive or negative given the current context, i.e. the current co-occurring terms.

### 3.1 Design Science Research

Design science research is one of the two prevailing paradigms in the area of information systems. While behavioral science, as its name indicates, concentrates on the analysis of human behavior and the behavior emerging when humans form organizations, the research in design science focuses on the creation and analysis of artifacts that further push the boundaries of human capabilities. These artifacts emerge from the stringent analysis of an aspect of information systems. Forming them requires the knowledge of already existing approaches as well as a search strategy, e.g. heuristical, that helps finding an artifact solving an existing problem. Hevner et al. (2004) emphasize that “technology and behavior are not dichotomous in an information system.” In other words, design science and behavioral science are not mutually exclusive - they collaboratively contribute to the formation of processes required in information systems.

Hevner et al. (2004) derive the relevance of design science research from the fact that the interface between business and information technology involves design processes. For instance, according to Henderson and Venkatraman (1993), creating an organizational infrastructure from a business strategy requires the availability of organizational design activities, as show in Figure 3.1. Furthermore, the infrastructure of an information system requires design processes inspired by the information technology strategy defined by the company.

The framework of design science research consists of three main pillars:

- **The environment**, consisting of the people, the organizations and the involved technology.
- **Research in information systems**, culminating in the design and evaluation of a useful artifact.

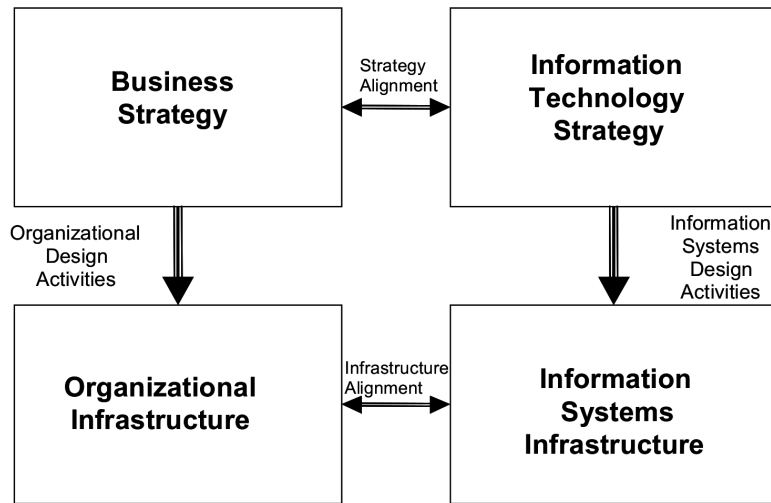


Figure 3.1: Design activities in an organizational environment (Henderson and Venkatraman, 1993).

- **The knowledge base**, where the knowledge gained during the process of designing and creating the artifact is fed back. Contributions broaden the knowledge of the scientific community.

These three pillars, also shown in Figure 3.2, are inter-connected. The *environment* articulates business needs, which are researched in the *IS research* pillar. To successfully accomplish the research, this pillar draws upon the information available in the *knowledge base*. It builds upon and integrates the foundations, e.g. theories, frameworks, and methodologies, e.g. data analysis, formalisms, or measures, already available. The outcome of successful research in design science requires the creation of an artifact previously unknown to the community. The artifact is further supposed to solve a problem that has not been solved by an existing approach before. After successfully finishing the study one goal is the re-integration of the artifact into the environment to prove its usefulness in the real world. Another goal is the integration of the gained knowledge into the knowledge base, thus expanding the existing knowledge and providing the means for the creation of further, even more sophisticated artifacts.

In order to identify design science research as such and to give assistance to research in the field, Hevner et al. (2004) provide a set of guidelines. The following section provides insight into these guidelines and connects the methods applied in this thesis to the guidelines.

### 3.2 Research Guidelines in Design Science

Following the principles of design sciences requires the compliance to a set of guidelines. These guidelines lead researchers in all steps of their work. They help finding relevant research goals and guide researchers in the implementation and evaluation of a prototype that tackles the goal and which might, in a later phase, find application in a practical environment. Hevner et al.

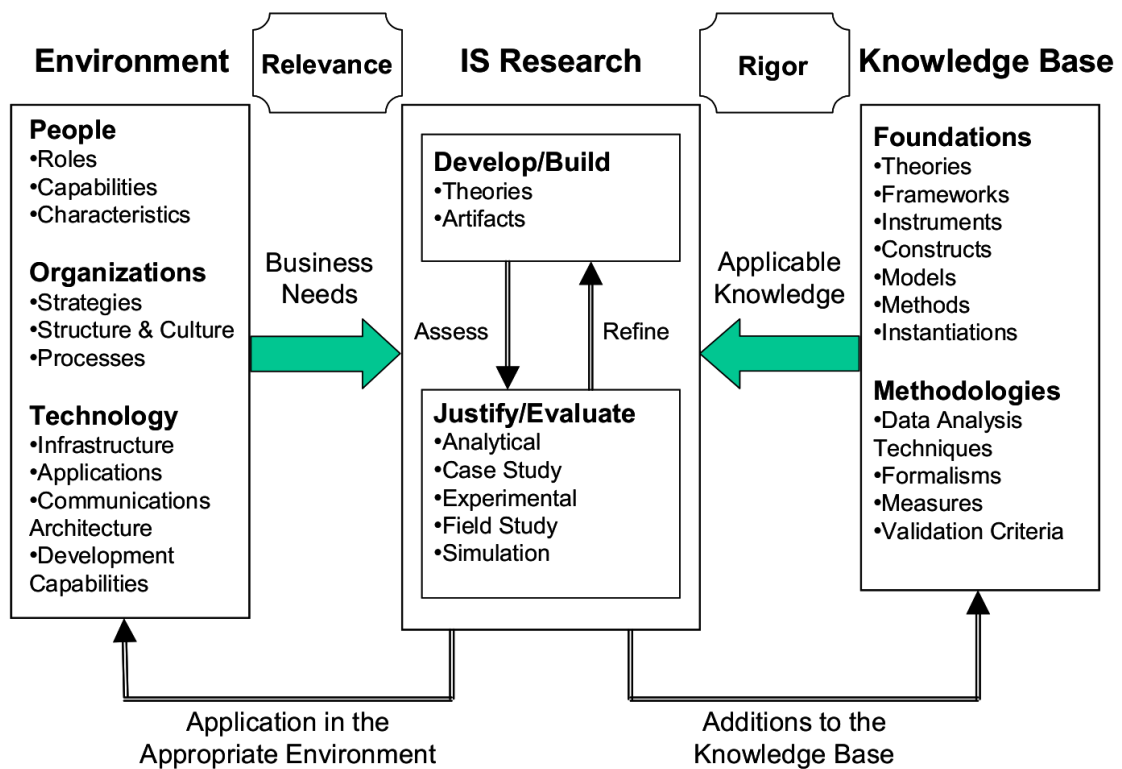


Figure 3.2: The framework of design science research with its three pillars.

(2004) define seven guidelines, outlined in the following sections. After defining each guideline its connection to the presented work will be illustrated and justified. They emphasize that the guidelines should be taken as such, i.e. as guidelines and not as a collection of obligatory laws. They suggest to use them flexibly and in a way so that they serve as an assistance to the judgement of researchers, but not as restrictions.

### Guideline 1 - Design as an Artifact

Applying design science as the scientific framework for research requires the creation of an artifact. In contrast to behavioral science, where the study of human behavior is the main focus, the area of design science concentrates on the creation of artifacts solving real-world information system problems. The creation, study, and evaluation of the artifact is the central matter of interest in design science. This artifact, consisting of “constructs, models, and methods”, solves a real-world problem.

The presented work results in the creation of a software prototype capable of solving a real-world problem inherent to opinion mining, i.e. the problem of polarity shifts in sentiment terms caused by the different contextual meaning. Analyzing and understanding the context helps to adapt a-priori sentiment values that are otherwise given in sentiment lexicons. This makes existing approaches more flexible and overcomes problems caused by static sentiment values.

For instance, it helps to decide whether *cool* in the “cool guy on the motorcycle” or “her face was cool and motionless when she broke up with me” depicts a positive or negative situation. The created artifact learns from real-world corpora, relieving the researcher from a cumbersome manual definition of this context knowledge.

## **Guideline 2 - Problem Relevance**

Design science requires the researcher to solve a problem without a solution thus far. This renders re-implementations for known problems as beyond design science and requires thorough study of the literature, which subsequently allows to isolate the research objective as unique. The presented work contributes to the so far unsolved problem of the impact of context in opinion mining. The created artifact analyses a given document collection and automatically identifies ambiguous sentiment terms based on statistical analysis. Subsequently, it identifies context terms helping to disambiguate, i.e. determine the optimal sentiment value, based on the given context. The approach is agnostic to the language of the document collection. It does not use syntactical rules that would tie its usage to a specific language but solely relies on statistical analysis. This facilitates its applicability across languages. The creation of the so-called contextualized lexicon, i.e. a lexicon that stores the ambiguous terms together with context terms and their probability values is only dependent on the availability of a training corpus in that language. Using corpora compiled from reviews, i.e. documents easily and cheaply downloadable from the Web, allows to quickly create the lexicon for a new language. Thus, a rapid improvement of sentiment lexicons in the given language becomes possible.

## **Guideline 3 - Design Evaluation**

Guidelines 1 and 2 focus on the creation of (i) an artifact that (ii) solves a problem without an existing solution. The third guideline of design science emphasizes the importance of an evaluation of the designed artifact. As the artifact might be used in a real use case it is necessary to confirm its efficacy. Moreover, its contribution to and integration into the scientific knowledge-base (see Guideline 4) require an appropriate evaluation to guarantee the relevance of the work.

The presented work fulfills this criterion by conducting 10-fold cross-validation on selected evaluation corpora and calculating evaluation parameters well-known in the field. 10-fold cross-validation is widely used to evaluate opinion mining systems. By shifting the samples in training and test corpora it virtually increases the size of the training and test space and reduces the risk of accidentally selecting an exceptionally beneficial or disadvantageous combination of training and test samples. Calculating statistical parameters allows an assessment of the overall efficacy of the approach and makes it comparable to the baseline. Applying a statistical significance test (i.e. Wilcoxon’s rank sum test) ensures that the obtained results are not a matter of accidental deviation. Without such a test, excluding mere luck as the origin of an improvement is not possible.

Recall, precision, and f-measure serve as the efficacy measurements. They are widely used in the research area. It is important to notice that none of these parameters should be shown in isolation, since they shed light on the problem from a different perspective. Please find more information about the parameters in Section 5.

Hevner et al. (2004) discuss five different lines of evaluation:

- **Observational:** Studying the artifact in the environment it is used, i.e. as an in-depth case study in a single environment or as a horizontal field study in multiple environments.
- **Analytical:** Studying certain characteristics of the artifact, e.g. its structure.
- **Experimental:** Running experiments to show the artifact's efficacy, either in controlled experiments or in a simulation.
- **Testing:** Observing the artifact while performing an action. The authors differentiate between black box testing, i.e. accomplishing stress tests on the artifact's interfaces, and white box testing, e.g. by executing parts of the artifact.
- **Descriptive:** Explaining the usefulness of the artifact, either from studying the knowledge base or by constructing scenarios.

The evaluation type of this work is *experimental*, i.e. the evaluation shows the efficacy of the proposed approach by running a set of experiments. Hevner et al. (2004) differentiate between a "controlled experiment", which means to "study [the] artifact in [a] controlled environment for qualities", and a "simulation", i.e. to "execute the artifact with artificial data". The design of the evaluation in this work is a simulation, although the prerequisite of "artificial" data is not entirely fulfilled. The data used for evaluation comes from movie, product, and holiday reviews and has been created by consumers of the products and services. Thus, it is not "artificial".

#### Guideline 4 - Research Contributions

Research in design-science demands clear contributions to the scientific knowledge base. The finished research work needs to deliver a contribution in one or more of the following areas:

- **The design artifact:** The created artifact itself can be the scientific contribution
- **Foundations:** Hevner et al. (2004) mention modeling formalisms, ontologies, algorithms, etc. as examples for foundations.
- **Methodologies:** Methodologies are new procedures created by the research work, e.g. new evaluation metrics.

The presented work contributes to the area of opinion mining in two ways. Firstly, it creates a fully-developed prototype tackling the problem of context in opinion mining, thus fulfilling the "design artifact" requirement. Integrating such an artifact into the software stack of web intelligence companies helps refining their modules for opinion mining. Secondly, it contributes to the "foundations" branch. The prototype produces so-called contextualized sentiment lexicons, helping researchers in this area to refine their sentiment lexicons.

The research contributions have been published to validate the relevance of the approach and receive feedback from the scientific community, resulting in the following track record:



- Gindl et al. (2010) presents the initial idea of contextualization. It describes the used Naïve Bayes approach and outlines an approach to overcome a common problem of machine learning techniques, i.e. the domain bias of the trained classifier. The author of the thesis was responsible to program the prototype and conducted the evaluation.
- Gindl (2010) builds upon previous insights (Gindl et al., 2010), and evaluates different strategies for cross-domain usage of contextualized lexicons.
- Weichselbraun et al. (2010) expands the set of evaluation corpora to further confirm the efficacy of the contextualization approach. The author of the thesis improved the previously implemented prototype and conducted the evaluation.
- Weichselbraun et al. (2011) documents the construction of a sentiment lexicon using a game-with-a-purpose (von Ahn, 2006) and expanding it with a bootstrapping technique. Games-with-a-purpose are designed to leverage human intelligence and to solve tasks that are beyond the capabilities of automated systems, e.g. image recognition tasks. The tasks are highly repetitive, causing fatigue in the persons working on them. The game setting circumvents this problem and maintains concentration and enthusiasm in the participants. A subsequent work describes the design decisions made to create the game-with-a-purpose (Scharl et al., 2012). The author implemented the bootstrapping module and developed an environment for the evaluation.
- Weichselbraun et al. (2013) and Weichselbraun et al. (2014) extend the initial work on contextualization. They suggest the usage of external resources such as WordNet or ConceptNet (Speer and Havasi, 2013) to further improve the contextualization approach. Contextualization, as described in by Gindl et al. (2010) is still strongly dependent on the quality of training corpora. Leveraging external resources such as WordNet or ConceptNet paves the way towards the integration of knowledge unavailable in the training corpora. Such a procedure improves the learning of the algorithm, providing it with abstract knowledge going beyond what it has already learned. The author focused on the improvement of the contextualization procedure.
- Gindl et al. (2013) covers opinion target extraction via grammatical patterns. The Stanford parser delivers dependency the dependency tree for a sentence. Applying grammatical patterns allows to identify the opinion target of a sentiment term. The author developed a graph-based prototype applying grammar rules and transmitting sentiment charges onto opinion targets.
- Clematide et al. (2012) describes the creation of a benchmark corpus. Benchmark corpora are essential for opinion mining. They provide the data to identify useful syntactical patterns or to train machine learners and allow for an evaluation of the created systems. Standardized corpora with broad acceptance in the research community, are the ideal resource to compare different algorithms with each-other. In opinion mining these resources are still sparse, especially in language other than English. The corpus created in this work consists of 270 sentences from the well-known DeWaC corpus Baroni et al. (2009) and is

annotated on the sentence-level, the word- and phrase-level and the expression-level. The author annotated the corpus.

- Ruppenhofer et al. (2014) summarizes the guidelines for a shared task on opinion holder and target extraction organized by the Interest Group on German Sentiment Analysis. Again, the author of the thesis annotated the data using specific annotation tools 2.3.

### **Guideline 5 - Research Rigor**

The research rigor guideline covers the necessity of the adherence to methodological, e.g. mathematical, procedures. However, Hevner et al. (2004) advise against a too strong focus on mathematical formalisms. When doing so, the resulting artifact might get impractical to use in real-world cases.

The artifact presented in this approach builds upon statistical insights of the used document collections. Furthermore, it employs the Naïve Bayes method to disambiguate ambiguous sentiment terms using context. The usage of these mathematical methods (outlined in Section 4) defines the research rigor in this work.

### **Guideline 6 - Design as a Search Process**

The search process in design-science research covers the identification of a proper solution for a given problem within the entire search space. Heuristic search helps paving the way towards a feasible and working solution. Formally, the *search* in design-science research is defined by its *means*, *ends*, and *laws*. *Means* are the methods employed to discover the solution. In the presented work, this is covered by finding appropriate preprocessing methods or setting proper threshold values for certain inclusion criteria. *Ends* represent the solution for the given problem. Finding a way of correctly leveraging context to disambiguate ambiguous sentiment terms is the goal, or the *end*, of this research work. Finally, *laws* are states or situations inherent to the environment that are unchangeable by the researcher. In opinion mining, researchers are confronted with the invariant characteristics of the underlying text, e.g. statistical contributions of terms or the existence of misspellings and grammar errors inherent to the nature of non-edited document types such as product and service reviews.

### **Guideline 7 - Communication of Research**

Communicating the outcome of design-science research is the last guideline. It requires communication with both a technology-oriented as well as a management-oriented audience. Hevner et al. (2004) define this as a “must”, rendering research solely presented to the scientific community outside of design-science. The presented work has been publicized in multiple scientific workshops, conferences, and journals. The management perspective is covered by its potential use in a business environment. webLyzard technology, a platform focusing on web intelligence, considers the inclusion of the artifact in its software stack.

To further support research communication the author of this thesis co-founded an interest group for sentiment analysis in the German language and has founded and co-organized a workshop in that area (PATHOS).

### **IGGSA - Interest Group on German Sentiment Analysis**

The Interest Group on German Sentiment Analysis is a community of researchers from Austria, Germany, and Switzerland, aiming to funnel the efforts in sentiment analysis for the German language. Founded in 2011, it has already taken efforts to contribute resources in the area: the MLSA (Multi-layered Sentiment Analysis) corpus, as described by Clematide et al. (2012), is a corpus compiled from German in total 270 sentences. These sentences, originating from the well-known DeWaC corpus (Baroni et al., 2009), are annotated on three levels: (i) the sentence-level, the most coarse-grained annotation, (ii) the word- and phrase-level, and (iii) the expression-level, beneficial for opinion holder and target extraction tasks.

More information on the activities of IGGSA is available here: <https://sites.google.com/site/iggsahome/home>, last accessed on 24 November 2014.

### **PATHOS - Practice And Theory of Opinion Mining and Sentiment Analysis**

PATHOS is a workshop that serves as a platform for the communication and knowledge exchange of researchers in opinion mining. The workshop is not tied to a specific language and welcomes contributions in all areas of opinion mining. PATHOS was first launched in 2012 collocated with KONVENS 2012, the “Konferenz zur Verarbeitung natürlicher Sprachen” (“Conference on Natural Language Processing”). In 2013, PATHOS was held in conjunction with the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2013). The proceedings of the workshop were published in a special issue of the Journal for Language Technology and Computational Linguistics, 29(1), 2014. In 2014, PATHOS was integrated into KONVENS as a track.

Summarizingly, design science research is a strong interface between the scientific and the business sector by providing the means to successfully conduct the research to create a cutting-edge technology artifact that is capable of being used in a real-world environment. It serves both the scientific community by extending its knowledge base and also business stakeholders, interested in expanding and improving their organizations’ software stacks with functionality tackling a so far unsolved problem. By providing a set of guidelines, Hevner et al. (2004) hand over a powerful tool to ensure the methodological rigor required for the creation of such an artifact.



## Methodology

*Never worry about theory as long as the machinery does what it's supposed to do.*

Robert A. Heinlein, *Waldo & Magic Inc.*

Opinionated language allows people to express how they feel about the environment they live in, e.g. their feelings about other people, their life situations, certain objects of their everyday life and other things. Without opinionated language, human communication is reduced to the mere exchange of facts. Factual information is merely descriptive, but does not imply any interpretation. Scientific language is an example where strong factual language prevails (or, at least, should prevail). As soon as people get emotive about a fact, they will draw on sentiment-carrying terms. These terms help others to interpret the feelings and emotions a person has towards a certain circumstance. The emotive character can be very diverse, ranging from expressions of rage or terror to admiration or amazement and other feelings as well as different levels of strength. For further information please refer to (Plutchik, 2001) or Section 2.5. Emotions such as “admiration” or “amazement” are generally considered as positive emotions. Having such a feeling means that the reason causing it is desirable. On the other hand, emotions such as “terror” or “rage” negatively upset a person and can eventually result in high physical distress. They are not desirable and leave the person in a state of discomfort.

The presented approach aims at solving a problem inherent in sentiment lexicons. Sentiment lexicons are collections of terms that are frequently used to express sentiment. For each term, the sentiment lexicon stores an a-priori sentiment value indicating its direction of polarity, i.e. positive or negative. While this approach is a valid first attempt, it neglects context-dependent shifts of the polarity of certain sentiment terms. These terms, in this thesis called “ambiguous terms”, can switch their polarity from one end of the scale to the other. One and the same term can have a positive polarity in a particular context, while it has a negative polarity in a different context. The following sections give an overview of the terms used throughout the remaining

part of the thesis (Section 4.2), outlines the approach (Section 4.3) and gives insight into the technical details (Sections 4.4 and 4.5), and explains the underlying theoretical model (Section 4.6). To start with, the next section gives a legitimation of this approach.

## 4.1 Legitimation

The digital representation of a sentiment term is its lexical transcription, i.e. the way the term is spelled, as well as a numerical value containing the polarity of the term. In case the term has negative polarity, it gets a negative value assigned and vice versa. To express different levels of polarity strength it is advisable to use a scale, e.g. from  $[-1., 1.]$  or  $[0., 1.]$ . Strong polar terms get values close or equal to  $-1. / 1.$  ( $0. / 1.$ , respectively), while weak polar terms center around  $-0.5 / 0.5$  ( $0.25 / 0.75$ ) or even come close to 0 (0.5). Table 4.1 shows example entries of the sentiment lexicon used in this thesis (for further information about the sentiment lexicon please refer to Section 4.6. “Joy”, as a strongly positive sentiment term, has a value of 1. assigned to it, while “anger” has a value of  $-1.$ . The weaker terms “trust” and “like” have 0.67 and 0.49 respectively. The values come from the General Inquirer (Stone, 1966), a term collection providing semantic information on categories such as religion or politics for approximately 12 000 terms. Sentiment information is one of the categories. The decimal numbers such as 0.67 for “trust” or 0.49 for “like” result from different meanings of the terms and the frequency distribution of this meaning in a reference corpus. For “like”, the distribution is as follows:

- 47% positive occurrences with the meaning *“To derive pleasure from, to find agreeable or congenial, to feel attracted to someone or something.”*
- 2% positive occurrences with the meaning *“Liking—the attraction to, pleasure in, or enjoyment of something or someone, fancy or inclination.”*
- 51% neutral occurrences with the meaning *“Having the same characteristics as, similar to, resembling, analogous to, in or after the manner of, for example, just as, such as.”*

The sentiment values in the used sentiment lexicon reflect the term distributions in meaning with either a positive or negative polarity and omit the neutral occurrences.

Term	Polarity value
joy	1.
anger	-1.
fear	-1.
hope	1.
trust	0.67
like	0.49
good	0.89

Table 4.1: Example entries in a sentiment lexicon.

The examples “trust” and “like” reveal a problem caused by the assignment of a-priori values to terms. Both of these terms are ambiguous in the sense that their sentiment value can switch in given circumstances. The following definitions and examples support the hypothesis that sentiment terms can switch their sentiment value. The definitions of the terms and the example sentences come from the Oxford dictionary<sup>1</sup>. For instance, “like”, used as a verb, expresses positive sentiment towards the concept a person likes:

“Find agreeable, enjoyable, or satisfactory,” e.g.

- “*all his classmates **liked** him,*”
- “*people who don’t **like** reading books*”

Used in a prepositional phrase “like” reveals a different behavior. Its semantic meaning changes from expressing positive sentiment to a comparative statement, eliminating its sentiment charge:

“Having the same characteristics or qualities as; similar to,” e.g.

- “*he used to have a car **like** mine,*” or
- “*they were **like** brothers*”

One strategy to tackle this problem is to differentiate between different part-of-speech tags and store multiple sentiment values for one and the same sentiment term, resulting in an expansion of the sentiment lexicon in Table 4.1 to the lexicon in Table 4.2.

Term	POS	Polarity value
like	VB (Verb, base form)	1.
like	IN (Preposition or subordinating conjunction)	0.

Table 4.2: Expanding a sentiment lexicon with POS tags in Penn Treebank style (Marcus et al., 1993).

However, such a strategy cannot solve all problems: for instance, a failure of the part-of-speech tagger can assign wrong POS tags to terms and consequently assign wrong polarity values to them. Colloquial language, as it is often used in forums or in micro-blogs, with its unclear language makes the accurate identification of POS tags a difficult task. The following examples further elaborate on this challenge. The examples show terms with ambiguous meaning and where a disambiguation based on POS tags will fail. Improving the accuracy of existing POS taggers cannot solve the problem completely. “Trust”, for example, has a positive meaning as a noun and a verb. According to the Oxford dictionary, “trust”, as a noun, denotes a

“Firm belief in the reliability, truth, or ability of someone or something,” e.g.

- “*relations have to be built on **trust**,*” or

<sup>1</sup>Oxford Dictionaries: <http://www.oxforddictionaries.com/>, last accessed on 24 November 2014.

- “they have been able to win the **trust** of the others”

The term’s polarity when used as a verb is also positive:

“Believe in the reliability, truth, or ability of,” e.g.

- “I should never have **trusted** her”, or
- “he can be **trusted** to carry out an impartial investigation”

However, “trust” has another meaning. It also denotes a particular type of companies, e.g.:

“An arrangement whereby a person (a trustee) holds property as its nominal owner for the good of one or more beneficiaries,” e.g.

- “a **trust** was set up”, or
- “the property is to be held in **trust** for his son”

An even more impressive example is the ambiguity of the term “good”. Intuitively, “good” always expresses positive sentiment. Its definition in the Oxford dictionary is:

“To be desired or approved of,” e.g.

- “It’s **good** that he’s back to his old self” or
- “a **good** quality of life”

Another definition is:

“Having the required qualities; of a high standard,” e.g.

- a **good** restaurant
- his marks are just not **good** enough

In both cases, “good”, used as an adjective, conveys positive sentiment. However, used as a noun, “good” reveals interesting behavior. In the following two senses it still conveys positive sentiment:

“That which is morally right; righteousness,” e.g.

- mysterious balance of **good** and evil

“Benefit or advantage to someone or something,” e.g.

- “he convinces his father to use his genius for the **good** of mankind” or
- “the preservation of old buildings matters because they contribute to the general public **good**”

There is a second meaning of “good” when used as a noun:



“Merchandise or possessions,” e.g.

- “*imports of luxury goods*” or
- “*stolen goods*”

The previous examples show sentiment terms that lost their sentiment charge. In other words, the context they were embedded in turned their sentiment charge to neutral. Complete shifts from one edge of the polarity scale to the other are also possible. For example, the term “killer”, intuitively a negative term, is capable of completely inverting its sentiment. The most intuitive meaning of “killer” is covered by this definition:

“A person, animal, or thing that kills,” e.g.

- “*a killer virus*”

Used in a different context, the term can invert its sentiment value:

“A formidable, impressive or difficult thing,” e.g.

- “*his new novel is a killer*”

A method limited to the usage of POS tags will fonder on these problems. Merely knowing that “good” is a noun does not allow any conclusions on whether the term has a positive meaning or not. To disambiguate its sentiment it is necessary to analyze its context. For instance, in the text snippet “*for the good of mankind*” the context term “mankind” serves as an indicator that “good” refers to the concept “righteousness”. Similarly, in “*the general public good*”, the terms “general” and “public” indicate the meaning “Benefit or advantage to someone or something”. In its neutral meaning the context terms “imports” and “luxury” serve as indicators for the neutral meaning of “good”. This shows that an in-depth analysis of the context helps to reveal the meaning of sentiment terms where other methods fail.

This problem can be approached in different ways. One option is to study the frequency distribution of the respective sentiment term in its context and count its occurrence in positive and negative context. Subsequently, the prevailing value becomes the a-priori value in the sentiment lexicon. Another option is to use the average sentiment value, i.e. count the number of positive and negative occurrences and divide it by the total number of occurrences. To make the a-priori values resistant to outliers other statistical measures such as median or mode can also be used. SentiWordNet, for instance, Esuli and Sebastiani (2006) tackles polarity shifts by storing different polarity scores for each term. They used a fine-grained method that assigns polarity scores to sentiment terms in different parts-of-speech. Figure 4.1 illustrates this strategy. Each sentiment term has a score (for each of its relevant parts-of-speech) positioned between positive, negative and objective (i.e. neutral).

A strategy such as using the average of positive and negative occurrences as a term’s sentiment value bears the risk of diminishing the polarity strength too strongly. For example, aggregating the sentiment value of a term with an equal distribution in positive and negative contexts to 0.5 results in turning the sentiment term into a neutral term. While the term might have had

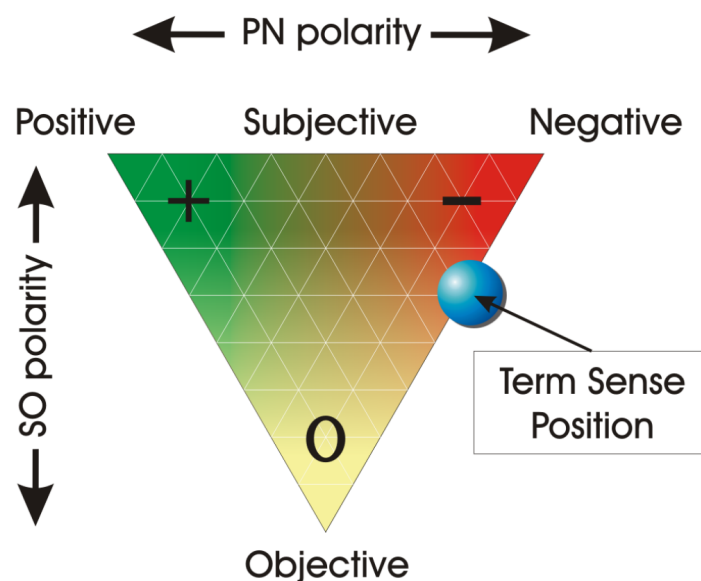


Figure 4.1: Representing ambiguities of polarities in SentiWordNet.

a very strong negative impact in negative contexts and a very positive impact in positive documents, its impact is distorted by such a procedure. Thus, this thesis proposes an approach that swaps sentiment values based on the context while retaining their strength. A sentiment term either has a positive value, i.e. 1., or a negative value, i.e. -1., but nothing in between. The system calculates the actual polarity of the term dynamically, i.e. it analyses its context and subsequently assigns it a positive or negative value.

## 4.2 Term Definition

Sentiment analysis uses its own vocabulary, whose terms sound familiar at the first glance but can be tricky to understand when thinking about them in more detail. The following section summarizes the important terms of this thesis.

**Sentiment term.** A sentiment term is a language token used in an opinionated statement. In this opinionated statement, the sentiment term itself conveys the type of the emotion (see Section 2.5) as well as its polarity and strength. Intuitive examples for sentiment terms are the terms “good” and “bad”.

**Sentiment lexicon.** This is a collection of known sentiment terms as well as information about their polarity and the strength. Sentiment lexicons usually contain several thousand sentiment terms. Sentiment lexicons can either be hand-crafted or compiled automatically. Hybrid approaches are also common and can be useful where manual annotation is too costly or many different lexicons are required, e.g. for different domains.

**Polarity.** Without regarding the more fine-grained emotion analysis, the polarity of a sentiment term can either be positive, negative or neutral.

**Strength.** The strength of a sentiment term is a measurement for the definiteness of a sentiment term. Intuitively, the term “excellent” is more positive than the term “good”. On the other hand, “hatred” is much more negative than just “dislike”.

**Sentiment value.** The sentiment value is a numerical value combining polarity and strength of a sentiment term. The range of this value is subject to algorithmical preferences but typically ranges in an interval between  $[-1, 1]$  or  $[0, 1]$ .

**Ambiguous term.** An ambiguous term is a sentiment term whose polarity changes depending on its context. Sentiment lexicons usually store a single value for a sentiment term and do not take context changes into account. An intuitive solution to this problem is to calculate an average sentiment value based on the term’s occurrences in positive and negative context. This value lies between the sentiment maxima. Yet, such a strategy reduces the power of a sentiment term. Instead of making it a term that is either positive or negative, the term becomes neither positive nor negative.

**Monosemous term.** In contrast to the sentiment value of an ambiguous term the value of a monosemous term stays constant and is unaffected by the context it is embedded in. Its distribution peak is in documents of the same polarity as their sentiment value.

**Context term.** Terms frequently co-occurring with ambiguous terms are called context terms. Each context term has a probability value assigned to it. This probability is the likeliness of the term to indicate the positive or negative usage of an ambiguous term. One and the same context term can have different probabilities for different ambiguous terms.

**Contextualized lexicon.** Expanding a sentiment lexicon with context information turns it into a contextualized lexicon. This kind of lexicon stores sentiment values for monosemous terms in the same way as a “traditional” sentiment lexicon does. In addition, for each ambiguous term it contains a collection of context terms as well as their probability to indicate the positivity or negativity of the ambiguous term. In a later phase, these probabilities determine the final sentiment value of the ambiguous term.

**Disambiguation.** Disambiguation collapses the polarity possibilities into a single value. It is the final selection of a sentiment value for an ambiguous term, which can be used for further calculations. After the disambiguation, the term has a fixed polarity, e.g. either positive or negative.

### 4.3 Overview

This thesis investigates the assumption that context plays a major role for sentiment analysis. As outlined in Section 4.1, the polarity value of some terms in a sentiment lexicon is not static but can change depending on the context. Thus, interpreting the context of such sentiment terms

helps determining their “actual” polarity. The system does this by creating a so-called “contextualized lexicon”. A contextualized lexicon expands a traditional sentiment lexicon with context knowledge. It differentiates between monosemous and ambiguous terms. Monosemous terms carry the same polarity in the prevailing number of cases and have a static polarity. Ambiguous terms can shift their polarity depending on their context. In the training phase, the system first identifies ambiguous terms in a sentiment lexicon. Subsequently, it gathers context knowledge by analyzing co-occurring terms. Adding this knowledge into the sentiment lexicon creates the contextualized lexicon. Using the contextualized lexicon in a sentiment analysis tool helps to identify ambiguous terms in a document. The subsequent disambiguation into positive and negative assigns a sentiment value to the term that is based on its context.

Summarizingly, the entire procedure consists of three major steps:

1. The identification of ambiguous terms
2. The identification of context terms
3. The application of the contextualized lexicon in a new sample

Figure 4.2 illustrates the procedure. In the first step, the system identifies ambiguous terms by analyzing the terms in a sentiment lexicon. Frequency distributions in a training corpus allow the conclusion whether a term is ambiguous or not.

In the second step, the system collects context terms for each ambiguous term. Context terms are all terms co-occurring with an ambiguous term in a document.

The third step is the application: the system gets an unknown document as input. It identifies monosemous and ambiguous term in the contextualized lexicon and uses the a-priori sentiment values of the lexicon for monosemous terms. For each ambiguous term, it analyses the co-occurring term in the unknown document and compares them with the context terms in the contextualized lexicon. By calculating the probability of the context to be either positive or negative it assigns the most likely sentiment value to the ambiguous term. After disambiguation, the systems summarizes all sentiment values and calculates an overall polarity.

## **Bag-of-Words**

The presented approach is a so-called “bag-of-words” approach, i.e. it is agnostic to the relations of the terms in the documents. The metaphor “bag-of-words” refers to any information retrieval approach disregarding the order of the terms in the documents - they resemble items that have been put into a bag and shaken thoroughly. They do not take into account relations between the terms, i.e. grammatical structures or semantic relatedness. Instead, the terms are considered as independent from each other. Such approaches are highly popular, although they lack the level of fine granularity that grammar-aware approaches have. Potential reasons are:

- **Ease of implementation:** In the simplest case a mere whitespace tokenizer is sufficient to turn a document into its bag-of-words representation, which strongly reduces to implement such an approach. More sophisticated tokenizers, capable of handling question marks, exclamation marks, colons, double quotes, etc. correctly, are available out of the

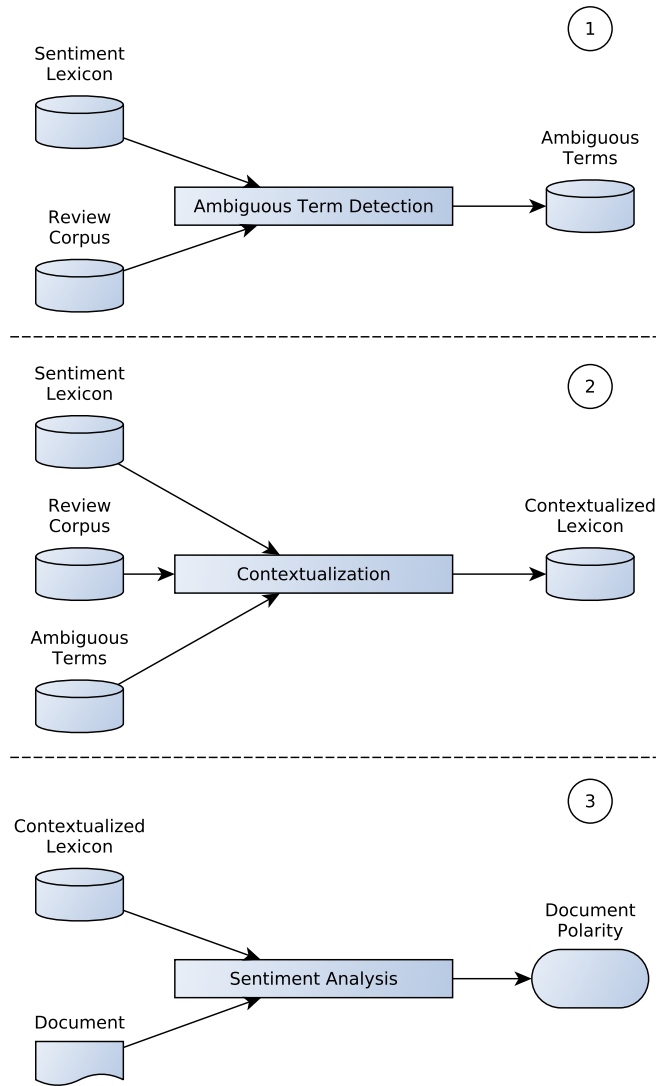


Figure 4.2: The contextualization procedure starting with 1) the identification of ambiguous terms, 2) continuing with the creation of the contextualized lexicon and 3) applying it to an unknown document.

box and integrated in natural language processing toolkits of many famous programming languages such as Python or Java. This makes them practical in situations where a quick proof-of-concept is desired.

- **High performance and scalability:** Since a bag-of-words approach omits grammatical relations and semantic relatedness, POS tagging a document or parsing it becomes unnecessary. This strongly reduces computational time and allows their usage in systems which are computation-intense and/or need to be scaled to a high level, such as search engines.
- **Acceptable efficacy:** Despite the naïve assumption that relations between the terms do not have an effect, bag-of-words approaches usually show an acceptable efficacy, which is tolerable given the two previously mentioned advantages.

The following sections describe each of these main steps in more detail.

## 4.4 Contextualization in Detail

### Identification of Ambiguous Terms

Monosemous sentiment terms have the peak of their frequency distribution in one particular sentiment class. For instance, if a monosemous term has a positive polarity assigned to it, it will occur in positive text snippets in the prevailing number of cases, but in negative ones only in rare cases. The rare number of occurrences in negative cases is due to exotic statements, e.g. ironic or sarcastic statements. For instance, “excellent” is a positive term. However, in the example “The camera I had ordered arrived with a broken lens. Excellent...” its sarcastic character results in a negative statement.

Given that the overall sentiment of a document in general stays the same it will, consequently, mainly occur in positive documents. The frequency distribution of this term has a peak in positive documents and a low level in negative documents. Ambiguous terms, on the other hand, are not so strongly tied to one specific polarity. They can express either positive or negative sentiment. Thus, they are likely to occur in positive and negative documents equally. Their frequency distribution does not have one single peak in one polarity class but is similar for both polarities. Figure 4.3 gives an exemplary frequency distribution for both monosemous and ambiguous sentiment terms.

The frequency distribution alone is an insufficient indicator for the ambiguity of a sentiment term. In case the lexicon contains a neutral term mistakenly this term’s frequency distribution will be similar to an ambiguous term’s distribution, i.e. an equal number of occurrences in both positive and negative documents. To overcome this problem and add a further cleaning step the system also uses a second parameter to judge ambiguity, which is the strength of deviation from neutrality. The documents used for the presented experiments are product and service reviews crawled from companies such as Amazon and TripAdvisor. Besides the text itself each review also has a rating, e.g. one to five stars as in the case of Amazon. To create a corpus with positive and negative documents from these reviews this work follows the general heuristic of considering reviews with a rating below three stars as negative and those with a rating above

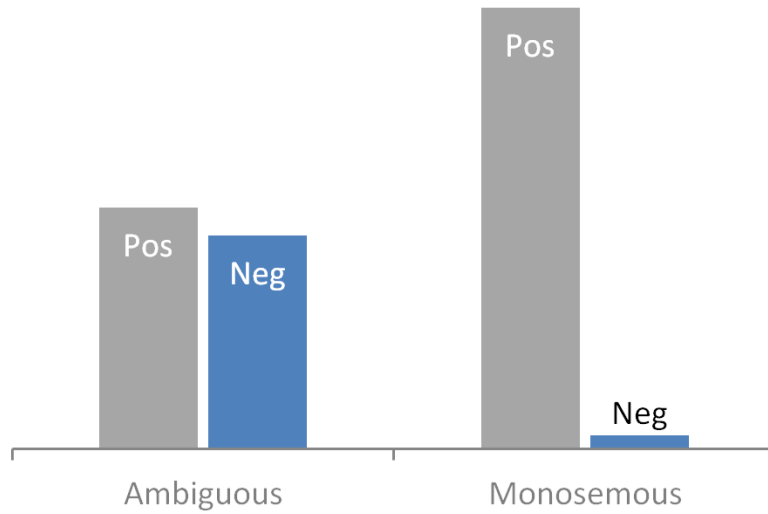


Figure 4.3: Exemplary frequency distributions of ambiguous and monosemous terms in positive and negative documents.

three stars as positive. Reviews with a rating of three stars are neutral. Given this experimental setup the average rating of the reviews a term occurs in indicates its ambiguity. Monosemous terms have peaks for either one/two star or four/five star ratings, whereas ambiguous terms are distributed equally among all ratings (see Figure 4.4).

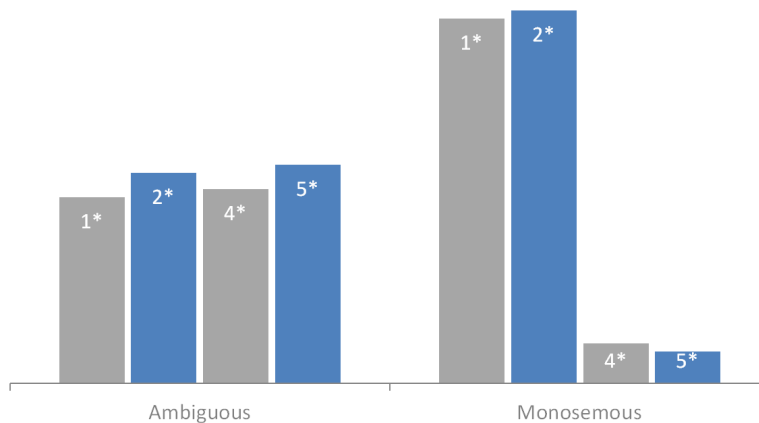


Figure 4.4: Exemplary frequency distributions of ambiguous and monosemous terms in reviews with ratings from one to five stars (three stars excluded).

An average close to neutrality indicates that the term is either neutral or ambiguous. To eventually distinguish between a neutral and an ambiguous terms its frequency deviation from the neutral value is crucial. The system considers terms with standard deviations above a certain threshold as ambiguous and those below as monosemous. Conducting a set of experiments

helped finding the most appropriate threshold values. Iteratively changing the threshold values and evaluating the method by comparing recall, precision, and f-measure has shown that the following values for standard deviation ( $\sigma$ ) and average frequency ( $\mu$ ) deliver the best results:

$$\sigma \geq 0.75 \quad \text{and} \quad (4.1)$$

$$\mu + \sigma \geq 0.25 \quad \text{and} \quad (4.2)$$

$$\mu - \sigma \leq -0.25 \quad (4.3)$$

The value of  $\sigma \geq 0.75$  ensures that the term is sufficiently ambiguous. The two thresholds  $\mu + \sigma \geq 0.25$  and  $\mu - \sigma \leq -0.25$  ensure that the term does not only deviate among the ratings in one polarity class but that it also sufficiently often occurs in reviews of the opposite polarity to be considered ambiguous.

Figure 4.5 summarizes the separate processes in the procedure. The system starts with counting how often the terms in a sentiment lexicon occur in the training corpus. It accomplishes a statistical analysis that serves as the basis for the categorization into ambiguous and monosemous. Applying a threshold filter to the statistical measures finally results in the compilation of a final list of ambiguous terms.

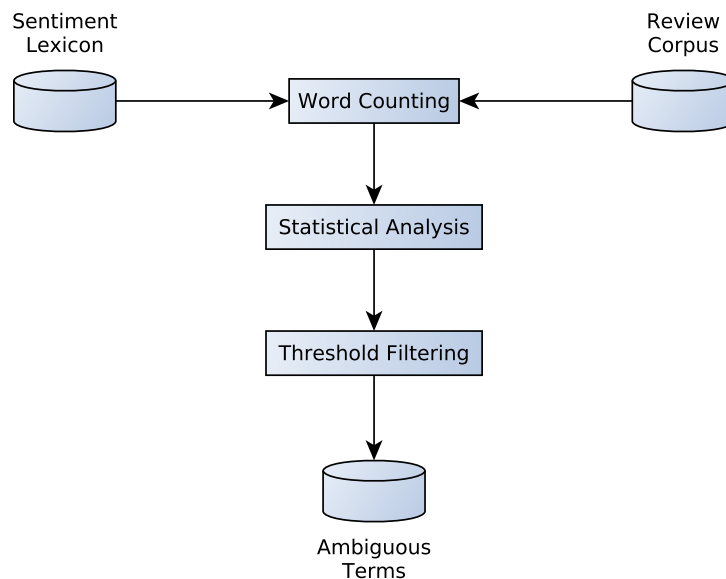


Figure 4.5: Identification of ambiguous terms in a review corpus.

After the identification of the ambiguous terms the system continues with the creation of the contextualized lexicon.



## Context Analysis

The next step in the creation of the contextualized lexicon is the context analysis. Collecting the co-occurrence frequencies of each ambiguous term/context term pair allows the calculation of a probability to estimate its likeliness to express positive/negative context. In other words, the system counts the number of co-occurrences in positive and negative context. Applying the Naïve Bayes technique on these co-occurrence frequencies results in probability values. This a-priori probability is an indicator of positive or negative polarity given that the ambiguous term and the context term co-occur in a document of unknown polarity.

After finishing the calculation of the probability values the system extends the existing sentiment lexicon with the probability information for each context term by adding the probability for positive co-occurrence:

$$p(C^+|c_i) \tag{4.4}$$

To avoid unnecessary redundancy the system does not store the probability for negative co-occurrence, since it can be easily calculated using  $1 - p(C^+|c_i)$ , i.e. subtracting the positive probability from 1. After adding context terms and their co-occurrence probabilities the sentiment lexicon has turned into a *contextualized sentiment lexicon*.

To illustrate the content of a contextualized lexicon, Table 4.3 gives an abstract glimpse into its content. Monosemous terms  $m_1, m_2, \dots, m_n$  have a static a-priori sentiment value, represented by  $ap(m_1)$ . Ambiguous terms  $a_1, a_2, \dots, a_n$  contain additional information, i.e. each context term  $c_i$  found in the training corpus as well as the probability for positivity  $p(C^+|c_i)$  in this training corpus.

Figure 4.6 summarizes the steps accomplished when creating a contextualized lexicon. The system uses the initial sentiment lexicon, the ambiguous terms identified in the previous step and the training corpus as input and accomplishes a statistical analysis. Applying the Naïve Bayes technique results in probability values for each ambiguous term/context term pair. This probability indicates the likeliness of this pair to occur in positive/negative context. The system eventually expands the original sentiment lexicon with context probabilities, resulting in a contextualized lexicon.

## Disambiguation of Ambiguous Terms

Given an unclassified document, the system at first identifies monosemous and ambiguous sentiment terms. For each ambiguous term it retrieves those context terms from the contextualized lexicon that are also present in the document. Based on the co-occurrence probabilities of the context terms it “disambiguates” the ambiguous term, i.e. it determines its probability given the context. The resulting sentiment value can either be identical to the one originally saved in the sentiment lexicon or the opposite of this value. The system uses the following formula for disambiguation:

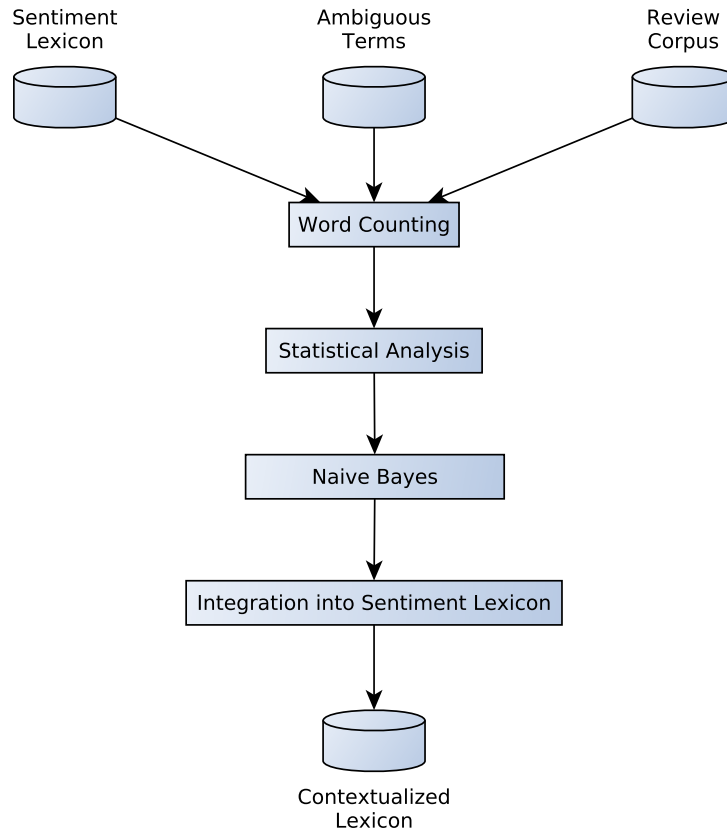


Figure 4.6: The contextualization procedure.

$$\mathbf{c} = \{c_1, \dots, c_n\} \quad (4.5)$$

$$p(C^+|\mathbf{c}) = \frac{p(C^+) \cdot \prod_{i=1}^n p(c_i|C^+)}{\prod_{i=1}^n p(c_i)} \quad (4.6)$$

Based on the sentiment values of monosemous and ambiguous terms the system calculates an aggregated overall value for the entire document. This overall value is a comparison of the number of positive and negative terms and uses the following formula:

$$sv_{total} = \sum_i^n sv(term_i) \quad (4.7)$$

Sentiment term	Context term	A-priori polarity/Probability
$m_1$		$ap(m_1)$
$m_2$		$ap(m_2)$
...		...
$m_n$		$ap(m_n)$
$a_1$	$c_1$	$p(C^+ c_1)$
$a_1$	$c_2$	$p(C^+ c_2)$
$a_1$	...	...
$a_1$	$c_n$	$p(C^+ c_n)$
$a_2$	$c_1$	$p(C^+ c_1)$
$a_2$	$c_2$	$p(C^+ c_2)$
$a_2$	...	...
$a_2$	$c_n$	$p(C^+ c_n)$

Table 4.3: Example entries in a contextualized lexicon.

## 4.5 Preprocessing

Raw textual data is often intermingled with disruptive artifacts, such as typos, or lacks particular information such as the data about used linguistic components. Preprocessing refers to all methods that clean up or enrich the original data. The following sections describe common preprocessing methods used in information retrieval and outline, how they are used in the artifact described in this thesis.

### Stop-word Filtering

Stop-word filtering is the exclusion of highly frequent terms, such as “the” or “a”, with only little valuable contribution to information retrieval tasks (Manning et al., 2009), such as the implementation of a search engine. Keeping stop-words results in an increase in storage requirements and computational time. Stop-word list are either hand-crafted or come from a statistical analysis of the underlying text corpus. A common heuristic is to count the term frequencies in the text corpus and cut out the  $k$  most frequent terms, e.g. the 50, 100, 200, etc. most frequent terms.

Despite the advantages of stop-word filtering, current search engines do store stop-words, i.e. they do not perform stop-word filtering. The assumption that stop-words are unnecessary is too naïve, since a search engine with active stop-word filtering cannot retrieve a phrase such as “to be or not to be”.

The artifact described in this thesis does not use stop-word filtering either. Instead, it uses a different approach to disregard insubstantial terms, as described in Section 4.6.

## Negation Detection

Negations are a crucial linguistic component in sentiment analysis, having the potential to completely alter the polarity of a sentiment term. A naïve yet widely used approach to handle negations is to invert the polarity of a sentiment term that is affected by a negation trigger:

The movie was *good*<sub>pos-pol</sub>. *Overall sentence polarity: 1.*

The movie was *not*<sub>negation-trigger</sub> *good*<sub>neg-pol</sub>. *Overall sentence polarity: -1.*

Discussion remains whether negation triggers really invert the polarity of a sentiment term or if they just annul its sentiment. In the latter statement the movie might just not have been good, which does not necessarily mean that it is bad. However, the common approach in the literature is to handle negations as sentiment shifters and not as diminishers (Polanyi and Zaenen, 2006). The system in this work adopts this approach. It inverts the polarity of sentiment terms if they are in the proximity of a negation trigger. The proximity is the three-term, right-hand proximity of the negation trigger. In other words, a negation trigger affects up to three terms after its position within a sentence. Negation triggers affect terms regardless of their part-of-speech, i.e. they affect adjectives, nouns, and verbs alike. A more extensive analysis of the exact behavior of negation triggers is beyond the scope of this work. Polanyi and Zaenen (2006) provide a good starting point to delve into this topic. Table 4.4 is the list of English negation triggers used in this work.

no	didn't
not	hasn't
never	haven't
without	hadn't
none	shouldn't
lack	wasn't
absence	won't
can't	wouldn't
couldn't	isn't
don't	aren't
doesn't	

Table 4.4: English negation triggers.

## Part-of-speech Tagging

Part-of-speech taggers assign a label to each token in a sentence, denoting it as nouns, adjectives, verbs, etc. The Penn Treebank tag-set (Marcus et al., 1993) offers 48 different labels, such as NN for a singular noun, NNS for plural nouns, or VBD for verbs in the past tense (see Table 4.5).

Tag	Linguistic particle	Tag	Linguistic particle
NN	Noun, singular or mass	PP\$	Possessive pronoun
NNS	Noun, plural	UH	Interjection
DT	Determiner	VB	Verb base form
JJ	Adjective	VBD	Verb past tense
JJR	Adjective, comparative	,	Comma
PRP	Personal pronoun	:	Colon, semi-colon

Table 4.5: Overview of the Penn Treebank tag-set.

According to Eugene Charniak, assigning a POS tag merely based on corpus frequencies, i.e. how often a term occurs as this part-of-speech, without any further knowledge, results in an accuracy of 90% (Charniak, 1997). Improving above this level requires the invocation of more sophisticated techniques. An important technique for POS tagging is the Viterbi algorithm (Viterbi, 1967), applied in a main class of POS taggers using Hidden Markov Models. Manning and Schütze (1999) provide an extensive introduction into this method. Systems using this technique are trained on pre-annotated corpora, e.g. the Brown corpus (Francis and Kucera, 1982), used to learn the probabilities of POS tag sequences. The trained POS tagger identifies POS tags according to their probability of occurring in a sequence. For instance, in the sequence  $t_1/DT, t_2/JJ, t_3/NN$ , the subsequent term  $t_4$  could have a 70% probability to be a verb, as in “the hardworking bee flies to the flower”, or a 30% probability to be a noun, as in “the hardworking bee keeper puts on his helmet.” The famous Brill tagger follows a different strategy by employing a rule-based approach (Brill, 1994).

The approach described in this thesis is agnostic to POS tags and does not use them for disambiguation. As discussed earlier, relinquishing POS tags eliminates the possibility to use them for disambiguation. On the other hand, it reduces computational time and retains the ability to apply the approach across languages.

The approach presented in this thesis uses the Naïve Bayes technique, a technique widely used in information retrieval. The following section gives an introduction into text classification in general, discusses available techniques and legitimates the application of Naïve Bayes.

## 4.6 Text Classification

Text classification is the process of assigning each document in a collection to a predefined class. Well-known examples are (Manning et al., 2009):

- **Topic detection:** The identification of the topic of an unknown document, e.g. for the retrieval of all topics related to a certain movie or a product.
- **Spam detection:** The filtering of unwanted messages, e.g. in an e-mail inbox. The term “spam” originates from the abundant usage of “spam” in a sketch of the comedy group Monty Python.<sup>2</sup>
- **Language detection:** The detection of a document’s language is necessary in cases where the software pipeline requires the invocation of language-specific resources, such as a POS tagger or dependency parser.
- **Sentiment classification:** In sentiment analysis, classifying a document means to assign a label *pos* or *neg*, which summarizes the overall sentiment value of the document. Sentiment classification can also refer to the task of assigning polarity values to phrases or single words.

Manning et al. (2009) point out that a rule-based approach to text classification, e.g. via the definition of sophisticated regular expressions, is possible but cumbersome. Such an approach usually requires the availability of extensive external knowledge, since creating sophisticated regular expressions is a highly complex task. This downside calls for automated methods for text classification. Finding those rules in an automated way reduces the necessity for manual input, thus lowering production costs and allowing an application on a large scale. Consequently, it is desirable to have an automatic method of learning to map documents to potential classes. Mathematically, this learning function maps each document  $d$  of the document space  $\mathbb{X}$  to a class  $c$  of the class space  $\mathbb{C}$  (Manning et al., 2009):

$$\gamma : \mathbb{X} \rightarrow \mathbb{C} \quad (4.8)$$

The learning function  $\gamma$  is learned by analyzing the features of pre-labeled training documents  $d \in \mathbb{D}$ .

In a wider context, text classification is a branch of machine learning, which is itself a sub-area of artificial intelligence. Machine learning is applied in various areas such as image recognition or language translation and is also widely used in sentiment analysis. A software implementation of a theoretical model of machine learning is called a “machine learner”, or simply a “classifier” (referring to its function of classifying entities into predefined classes). Machine learning involves a training phase, where the classifier learns the characteristics of the given collection of entities in regards of their class label. This fact, i.e. the learning of characteristics given externally defined labels is also referred to as “supervised learning”. Supervised learning is a learning procedure where the classifier can refer to pre-given labels, usually manually defined and assigned to the entities of the collection. In contrast, “unsupervised learning” does not require the definition of these classes. The unsupervised learner attempts to find potential classes itself. An example for unsupervised learning is clustering.

---

<sup>2</sup>For further information about the sketch please refer to Wikipedia: [http://en.wikipedia.org/wiki/Spam\\_\(Monty\\_Python\)](http://en.wikipedia.org/wiki/Spam_(Monty_Python)), last accessed on 24 November 2014.

After the training phase, the classifier is ready to be applied to unlabeled, new documents. Based on the characteristics the classifier has learned in the training phase it now attempts to find the most appropriate label for a given unknown entity.

Many classifiers are binary, i.e. they only distinguish between two different labels. Examples for binary classifiers are the Naïve Bayes technique or Support Vector Machines. Other classifiers such as Perceptrons or Decision Trees can distinguish between several classes. Combining several binary classifiers allows for the categorization into multiple labels, which is referred to as ensemble learning.

Evaluating the efficacy of classifiers is crucial before they can be applied in a real-world scenario. Standardized parameters such as recall (a measurement of completion), precision (a measurement of exactness), f-measure (the harmonic mean of recall and precision) and accuracy (another measurement of exactness) help assessing their efficacy. Section 5.2 further explains the statistical parameters used in this thesis.

The following section describes the so-called feature selection, a procedure used to reduce the input space of a classifier.

## Feature Selection

Feature selection is the procedure of choosing items in the data space most relevant for the classification task. In information retrieval, the most relevant features are terms with a high linguistic impact, usually nouns, adjectives, or adverbs. Determiners or conjunction words (also called stop-words) are less relevant. Feature selection is a delicate procedure and requires ample attention because of domain differences. For example, emoticons such as “:-)”, “:-(”, or “<3” are highly relevant in sentiment analysis. Their heavy usage in micro-blogs such as Twitter makes them an indispensable component for a sentiment analysis system. Micro-blogs require special attention: the widely used hashtag “#” adds additional information, e.g. to denote “#sarcasm” or “#irony”. Both sarcasm and irony are stylistic elements that are still challenging in the sentiment analysis area.

The purpose of feature selection is two-fold: on the one hand, it reduces the feature vector, which reduces the time required to train the classifier. On the other hand, it also reduces noise in the feature space and helps improving the classification accuracy. Common feature selection methods are:

- **Mutual information:** measures, how well a feature performs as an indicator for a particular class. For instance, a term occurring in documents of only one particular class is a strong indicator for that class, whereas a term equally spread amongst documents of all classes is a weak indicator.
- $\chi^2$  **feature selection:** detects features with high inter-dependence, e.g. features with a high chance of co-occurrence and thus a high likeliness to have a crucial impact.
- **Frequency-based feature selection:** discard terms with rare occurrence.

The approach presented in this thesis employs a sophisticated statistical filtering method that selects context terms based on their frequency and impact, as described in the following section.

## Filtering Insubstantial Terms

When creating the contextualized lexicon the system adds each term co-occurring with an ambiguous term, and does not perform any stop-word filtering. This makes the system completely flexible, allowing it to be employed in other languages without any further manual input, i.e. it is not necessary to provide a list of stop-words for a new language. This gain in flexibility comes at the price of an increased introduction of garbage terms for the calculation of the final sentiment value. To overcome this problem, the system uses a selection strategy to filter out garbage terms during the disambiguation process. It only uses the  $k$  context terms with the strongest probability for positive and negative. Experiments have shown that  $k = 10$  is a good threshold for the domains of this work.

The following section describes algorithms well-known in information retrieval (Manning et al., 2009).

## Support Vector Machines

Given a number of data points, Support Vector Machines try to identify a hyper-plane in an  $n$ -dimensional space that most properly separates the data points. The kernel trick helps to transform the input data into the  $n$ -dimensional space so that a separating hyper-plane can be found more easily.

The hyper-plane is constructed in a way so that the margin between itself and the closest data points is maximized. These data points are also called the support vectors. Figure 4.7 shows an illustration of the working principle.

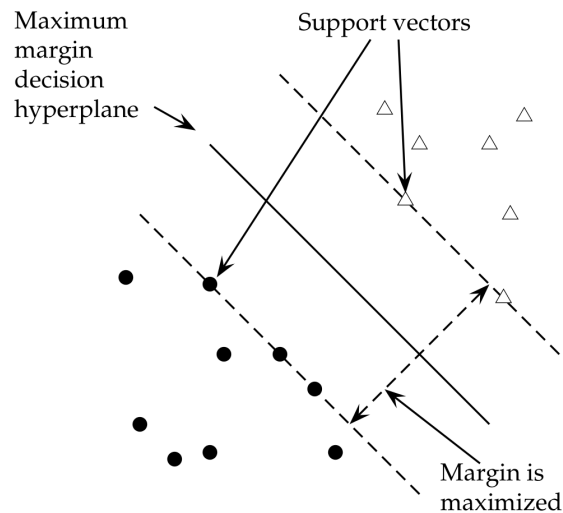


Figure 4.7: A hyper-plane separates the data points (Manning et al., 2009).



## Nearest Neighbor

This classification algorithm assigns each data sample to the class containing the samples with the smallest distance to itself. In other words: during the training phase, the algorithm merely stores the feature vectors of the data samples and the class they belong to. During the classification phase, the algorithm compares the features of each training sample to the new sample. Subsequently, the new sample gets the sample as the training samples with the smallest distance has. “Distance” refers to the vector space distance, common methods for its calculation are:

- **Euclidean:**  $\sqrt{(x - p)^2}$
- **Euclidean squared:**  $(x - p)^2$
- **City-block:**  $abs(x - p)$
- **Chebyshev:**  $max(|x - p|)$

The algorithm classifies a new data sample equal to the class of the majority of its nearest neighbors. For  $k = 3$ , it assigns the new sample to the class of at least two of its nearest neighbors. With a  $k = 5$ , it gets the class of at least three of its nearest neighbors. Figure 4.8 illustrates the behavior of the algorithm. The grey diamond, representing a new data sample, will be assigned to the class of green circles for  $k = 3$ , visualized by the inner circle with the solid line. For  $k = 5$  it will be classified as a red square, contained in the outer circle with the dashed line.

A single binary classifier separates data into classes  $A$  and  $B$  but fonders on the classification of classes  $A$ ,  $B$ , and  $C$ . Two pipelined binary classifiers master the problem. The first one classifies the data into  $A$  and  $\neg A$ . The second gets the  $\neg A$  data as input, which is either  $B$  and  $C$ , and classifies accordingly. Further classes require more binary classifiers, resulting in a complex chain of classifiers. The  $k$  nearest neighbor algorithm is an  $n$ -ary classifier, i.e. it supports the assignment of more then two labels without pipelining several classifiers.

## Naïve Bayes

Naïve Bayes is a widely used classifier thanks to characteristics such as its high accuracy and computational velocity. The attribute “naïve” reflects its neglection of relations between the features of a data sample. It assumes that the features are independent from each other, which is not true for information retrieval because of grammatical structures or semantic relatedness. Despite this, the technique still provides highly accurate results for information retrieval tasks.

The Naïve Bayes algorithm assigns a document  $d$  to a class  $c$  depending on the probability of each term  $t_k \in d$  to occur in the given class. During the training phase, the algorithm calculates  $P(t_k|c)$ , i.e. the probability of term  $k$  to occur in class  $c$ . When classifying a new document  $d$ , the algorithm uses the probabilities of all the terms in  $d$  and returns the most likely class for  $d$ . Mathematically, this procedure can be summarized in the following equation:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (4.9)$$

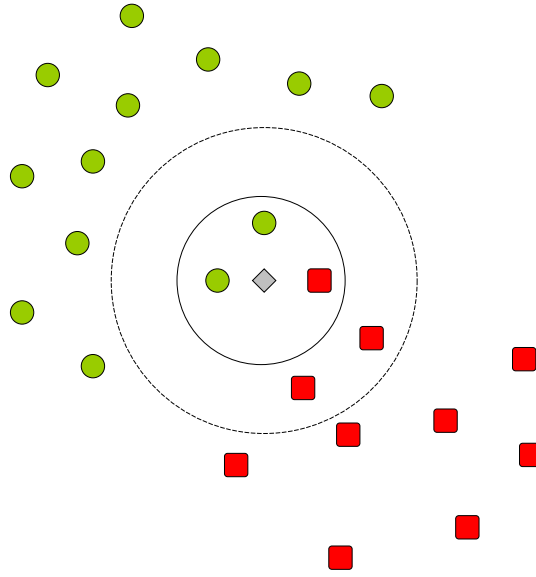


Figure 4.8: Different  $k$  values result in differing classifications, requiring knowledge about the data to choose the most promising value.

In cases where the terms in the document do not deliver enough evidence for the classification to be completed, i.e.  $P(t_k|c)$  does not give a convincing result, the algorithm uses the a-priori polarity  $P(c)$  to assign it to a class.

The Naïve Bayes technique follows two major approaches:

- **Multinomial Naïve Bayes:** Calculates the probabilities based on the number of occurrences of a term in a document, i.e. multiple occurrences of one and the same term in a document change the probabilities.
- **Bernoulli Naïve Bayes:** Multiple occurrences of a term in document are only counted once, i.e. a terms is considered as either in the document or not. The efficacy of this method decreases with increasing document length.

Both the multinomial and the Bernoulli model have the same linear time complexity. The Naïve Bayes technique encounters a mathematical problem when the new data, or test data, contains terms without an occurrence in the training data. In this case,  $P(t_k|c)$  results in a value of 0. Such a 0 value causes the product of all term probabilities  $\prod_{1 \leq k \leq n_d} P(t_k|c)$  to be 0 as well. Adding 1 to each term count helps to avoid this problem. This strategy is called “Laplace smoothing” and can be expressed with the following formula:

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B} \quad (4.10)$$

The term  $B$  denotes the vocabulary of the document collection, i.e. all terms used for the classification task, with  $B = |V|$ .

The Naïve Bayes technique provides the mathematical basis for the artifact implemented in this thesis. Its ease of implementation, computational performance as well as its robustness as a all-round classifier were the main reasons to use this classifier.

### Description of the Used Sentiment Lexicon

The sentiment lexicon used in this thesis originates from the opinionated terms available in the General Inquirer (Stone, 1966). The General Inquirer is a semantic lexicon, containing knowledge about a diverse set of information, such as political (the *polit* tag), religious (the *relig* tag), or even landscapes (the tags *aquatic*, *land*, and *sky*) for 11 985 terms. Among the available categories there are also the categories *pos* and *neg*, assigning the polarity of a term. We found 1 711 terms with the label *pos* and 2 021 with the label *neg*. These 3 732 terms served as the seed terms for the sentiment lexicon used in this work. The terms in the General Inquirer are in their infinitive form, i.e. plurals (e.g. “war”, “wars”) and flexions (“love”, “loves”, “loved”), are missing. A reverse lemmatization procedure added suffixes to the infinitives, extending the seed lexicon to 8 276 terms in total, with 3 195 positive and 5 072 negative terms. Table 4.6 contains ten examples terms of the lexicon.

Term	Polarity
abyss	-1
blame	-1
blessed	1
holy	1
lack	-1
misfortune	-1
obscure	-1
perfect	1
quit	-1
raid	-1

Table 4.6: Example terms in the used sentiment lexicon.

Using the General Inquirer for sentiment analysis has a long history and is well-established in the literature (Denecke, 2008; Esuli and Sebastiani, 2006; Thelwall and Buckley, 2013; Turney and Littman, 2002).



# Evaluation

*Truth, she thought. As terrible as death. But harder to find.*

Philip K. Dick, *The Man in the High Castle*

Hevner et al. (2004, p. 85) emphasize that “evaluation is a crucial component of the research process” in the sense that it allows to estimate the efficacy of a newly implemented artifact. In other words, it allows to assess if the research goal was met. A formal evaluation requires the adherence to well-defined standards in this research area. The work presented in this thesis is strongly connected to information retrieval, which necessitates an evaluation common in this area.

Evaluating information retrieval tasks involves the usage of a well-defined procedure. Annotated corpora serve as the basis for the evaluation (see Section 5.4). Training and testing the artifact using cross-validation, i.e. a controlled way of rotating training and test samples, artificially expands the usually sparse data and allows testing the artifact on more unknown data samples (Section 5.3). The application of statistical parameters widely used in information retrieval, i.e. “recall”, “precision”, and “f-measure”, allow to have a detailed look at the characteristics of the implemented artifact. Each parameter gives different insights, which makes it necessary to combine them to understand the entire picture (see Section 5.2). For more details on the used evaluation concepts and statistical parameters as well as a critical discussion please refer to Manning et al. (2009).

## 5.1 The Baseline

The baseline is a benchmark for comparison with the implemented system. In many cases it is a state-of-the-art algorithm or a system currently used. During the evaluation, comparing the results from the new system with this baseline shows whether the new system is promising.

In the present work, the baseline is a strategy commonly used in sentiment analysis, a so-called keyword lookup. This approach maps terms in a document to their sentiment values in a sentiment lexicon. By calculating a ratio between positive and negative terms the system aggregates an overall value for the document.

$$sv_{total} = \frac{n_{pos}}{n_{pos} + n_{neg}} \quad (5.1)$$

The above formula does not differentiate between strong and weak polarity but treats it equally. To overcome this problem, the baseline uses the following formula (equal to Equation 4.7):

$$sv_{total} = \sum_i^n sv(term_i)$$

The algorithm follows the negation detection strategy proposed by Polanyi and Zaenen (2006) and inverts the polarity of sentiment terms affected by a negation trigger. A negation trigger affects a sentiment term if the sentiment term appears within a frame of three tokens following the negation trigger, e.g. “This movie is **not bad**” or “This is not a bad movie”. The following formula summarizes this strategy:

$$sv_{total} = \sum_{t_i \in doc} n(t_{i-k})[s(t_i)] \quad (5.2)$$

where

$$1 \geq k \geq 3$$

Keyword lookup algorithms are popular because of their ease of implementation and low computational costs. Lightweight implementations merely involve the lookup in a dictionary data structure as well as a summation function to aggregate the overall values. Consequently, such an approach is favorable in cases with high data throughput.

## 5.2 Efficacy Measurements in Information Retrieval

The crucial concept for the evaluation of an information retrieval system is the notion of “relevance”. Relevance applies to user queries and is given when a system retrieves a document that fulfills the request formulated in the query. For instance, retrieving the information about the fourth planet of the solar system for the query “mars AND red AND planet” is a relevant search result. On the other hand, returning information about Mars, the Roman god of war, is irrelevant in that context. However, it is often difficult to ultimately answer the question whether a document is relevant for a query or not. The mentioned Mars query might, for example, also refer to the movie “Red Planet Mars”<sup>1</sup>. In the end, only the person formulating the query can assess whether it has been answered satisfactorily or not.

<sup>1</sup><http://www.imdb.com/title/tt0045073/>, last accessed on 24 November 2014.

The development of formulas to calculate the efficacy requires variables for counting the results. A document retrieved for the query that is in fact relevant is called a “true positive” result. In case the system mistakenly retrieves a document as relevant this is called a “false positive”. All potential combinations result in the four variables “true positive”, “false positive”, “true negative”, and “false negative”, as described below, :

- True positive: A retrieved document that is relevant.
- False positive: A retrieved document that is actually irrelevant.
- True negative: A rejected document that is irrelevant.
- False negative: A rejected document that is actually relevant.

The following sections describe statistical parameters well-known in information retrieval as well as the significance test used in this thesis.

## Recall

Recall measures how many of the desired samples have been successfully detected by the system. A high recall means that only a small number of samples was left undetected. In the context of text mining, a high recall means that the system was able to detect many documents relevant for the specific task. The mathematical representation is:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (5.3)$$

The following formula expresses the same concept but uses the aforementioned notation with true/false positives/negatives:

$$\frac{tp}{tp + fn} \quad (5.4)$$

## Precision

Precision measures the number of wrong guesses of a system, i.e. how many samples have been mistakenly classified as relevant ones. In text mining, precision measures how many documents have been mistakenly considered as belonging to a desired class, e.g. how many documents with an actually negative sentiment were classified as having positive sentiment and vice versa. The following formula calculates precision:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (5.5)$$

The alternative expression using true/false positive/negative is:

$$\frac{tp}{tp + fp} \quad (5.6)$$

## F-Measure

Exclusively measuring recall without measuring precision or vice versa is dangerous. For instance, to maximize recall the system simply needs to consider every data sample as belonging to the desired class. In that case, precision is likely to drop to a small value. Without knowing precision, misinterpreting the high recall as a radical improvement of the system becomes likely. Another parameter, the so-called f-measure, combines recall and precision to avoid misinterpretation. F-measure is the normalized, harmonic mean of recall and precision:

$$\text{f-measure} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (5.7)$$

A highly common value for  $\beta$  is 1, also called the  $F_1$ -measure:

$$f_1\text{-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (5.8)$$

Usually it is advisable to use the three parameters in combination as it allows to look at the efficacy of the classification task from different perspectives.

## Wilcoxon's Rank Sum Test

The measurements discussed in the previous section deliver relevant information about gains or losses in performance when a new algorithm or system is implemented. However, performance changes can still be a matter of mere luck. Consequently, it is necessary to rule out or minimize the risk that chance influences the result. This can be accomplished by using so-called significance tests. Significance tests are a well-known procedure in statistics. They are employed to reject the null hypothesis, i.e. the assumption that a result change is caused by chance.

A well-known significance test is the so-called  $\chi^2$  test, which finds application in medical or psychological experiments. However, the nature of the data examined in this thesis makes a different significance test more useful. The examined data are recall, precision, and f-measure values from ten rounds of cross-validation (see Section 5.3). A higher value of recall/precision/f-measure equals a better result or *rank*, implying a potential ranking order of the values. The  $\chi^2$  test is agnostic to ranks, which can cause a misinterpretation of the results.

To avoid this problem, the significance test used in this work is Wilcoxon's rank sum test, introduced to compare the efficacy of two treatments (Wilcoxon, 1945). The outcome of each round of cross-validation gets a rank. Comparing the rank sums of each procedure results in a significance value. In cases where this significance value is below a particular threshold the result is "significant", which means it is unlikely that the it was caused by chance. This work considers significance levels below 0.05 as significant.

## 5.3 Cross-validation

Evaluating a machine learner requires the availability of a test sample that has not been used for training. Using training data for the evaluation introduces the risk of bias and data overfitting. In other words, it remains unclear if the classifier only works on the data it has been



trained on and fails on new data or if it is flexible enough to reliably classify data samples with characteristics differing from the training input. Thus, the available annotated data needs to be split into a training and test partition. A common technique is to dedicate 90% of the data for training and the remaining 10% for testing. Such a strategy is legitimate but has the unwanted side effect of reducing testing data to a small amount. A way to overcome this problem is to shift training and test data. In a first round, the first 10% are used for testing and the last 90% for training. In the second round, the second 10% are used for testing while the remaining 90% are for training. Accomplishing this procedure ten times in total means that all data samples have been used for testing at least once (see Figure 5.1). This means that the size of the test data is ten times higher with this strategy than without it.

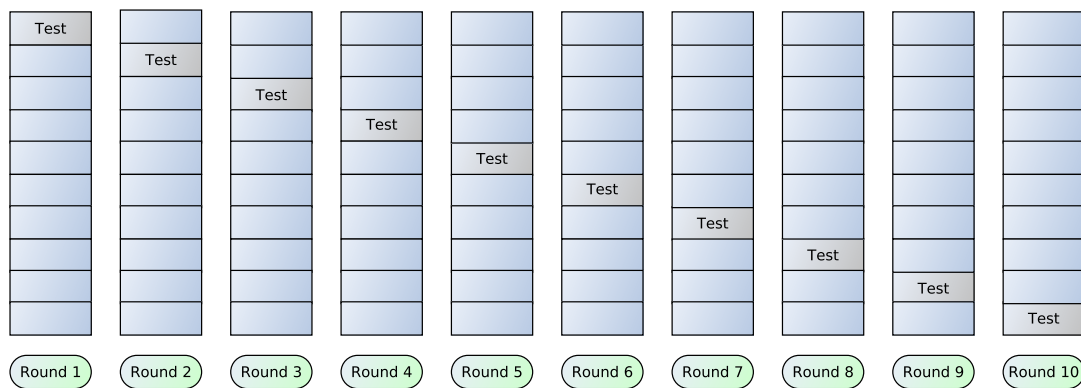


Figure 5.1: The test partition moves until every part of the document collection has served as test partition at least once.

The described procedure is called “10-fold cross-validation”. Splitting the data in an intelligent way and rotating the test and training samples meaningfully increases the number of samples artificially and reduces the risk of accidentally splitting the data in a “fortunate” or “unfortunate” way.

Cross-validation is widely used in information retrieval. 10-fold cross-validation is highly common, but many approaches also use 3-fold cross-validation. Here the ratio between training and test data is 2 : 1, resulting in three rounds in total. Having as many rounds as data samples, i.e. in each round there is only one test sample, is also called “leave-one-out cross-validation”.

## 5.4 The Evaluation Datasets

A common approach to create evaluation corpora in sentiment analysis is to use reviews downloaded from the World Wide Web (Klinger and Cimiano, 2014; Liu, 2012; Pang et al., 2002; Turney, 2002; Waltinger, 2010). Reviews exist for a plethora of topics: Amazon.com hoards extensive amounts of reviews for their diverse assortment of products; TripAdvisor.com provides reviews on hotels and hospitality services; IMDb.com is a database for movie reviews. Other

websites such as Epinions.com completely focus on providing reviews without giving any other service.

An evaluation dataset for sentiment analysis has two requirements: it needs (i) a written text about a certain topic and (ii) a label indicating whether the text has an opinion and, if so, the polarity of that opinion. To compile such a corpus it is possible to read through documents and subsequently apply a label for the read text snippets. In order to avoid mistakes it is necessary that several annotators read the same texts and also label these texts. Calculating an inter-annotator agreement ensures that the texts are not too ambiguous and help discovering systematic mistakes, e.g. intentional misclassifications by an annotator.

Such a strategy is promising but also cumbersome: to create a corpus with a size sufficient for meaningful evaluation, the collection of hundreds or thousands of samples is necessary. While this is feasible for single sentences it becomes a burden when trying to collect entire documents. Reading one document requires a particular amount of time. Side effects such as fatigue, loss of concentration, or simple boredom, further complicate the task. Moreover, prior to starting the annotation, it is necessary to set up guidelines that help the annotators to completely understand the task. Occasionally it is necessary to have several rounds of discussion before the guidelines are set up.

A shortcut to the creation of an evaluation dataset lies in crowd-sourcing. Crowd-sourcing exploits the “Wisdom of the Crowds” (Surowiecki, 2005) by exposing tasks to the people on the Web. Common tools are Amazon’s Mechanical Turk or CrowdFlower. These platforms allow to set up tasks requiring human intelligence to be solved. The tasks usually require a low skill level, but cannot yet be reliably solved by computers either, e.g. image recognition tasks. Another option to set up a crowd-sourcing task are “Games with a purpose”, where the task is presented in a way that the users believe they are playing a game. The inventor of this concept is Luis von Ahn, who coined the term “human computation” and invented the “Extra Sensory Perception Game”, the first game with a purpose aiming at image recognition.

Crowd-sourcing creates a new set of challenges. Due to the lack of personal communication it gets significantly harder to formulate the task specifications. The humans working on the tasks can hardly ask the employer in case they do not understand the specification. Furthermore, crowd-sourcing is vulnerable to cheating and requires special strategies to avoid a pollution of the data caused by cheating (see Section 2.2).

To overcome the mentioned problems another strategy is promising: since the rise of the Web 2.0 it has become a common pastime for people to share their opinions on a diverse set of topics via the World Wide Web. The simplification of Web editing tools enabled technically unskilled persons to set up their personal Web space and share their thoughts with the world. One way are platforms to create reviews for different topics. For instance, after the purchase of an item on Amazon the customer is encouraged to write a review about the purchased product. This possibility gives customers the feeling to be able to interact with an otherwise intimidating and invisible big company and also serves as quality measurement. As a side effect, the authors of these reviews also create a knowledge base of utmost relevance for researchers in sentiment analysis.

Each review is a written text, with the innate intention to express an opinion. Furthermore, the authors usually summarize the opinion they just expressed in the text with a label indicating

the polarity. Usually, these labels are star ratings or simple positive/negative labels. Thus, using reviews in an evaluation corpus eliminates the need for manually reading the text and classifying them, because they have already been classified by the author. Moreover, the risk of misclassification is reduced: the authors exactly know the feelings they have about a certain product and can easily summarize them in an overall label.

Thus, collecting a corpus of reviews is advantageous in two ways. On the one hand it is guaranteed that the text actually expresses an opinion, and this opinion is summarized and easily extractable from the rating. On the other hand, reviews can be downloaded easily on a large scale. Setting up a review crawler can be accomplished using out-of-the box toolkits. In many cases the websites provides an API for easy access themselves.

A potential downside of this approach is the lack of corpus cleanliness. While curated texts are usually of high quality in terms of grammar and orthography, this does not apply to the texts found on the Web. People tend to neglect orthographic and grammatical rules. Furthermore, the Web develops its own characteristics. Abbreviations are heavily used, e.g. 'I luv u' for 'I love you', and special character combinations, so-called emoticons, replace facial expression, e.g. ':-)' for a smile.

These characteristics create further challenges for the researchers, but the advantages outweigh the detriments. Automatic spell checkers are easily available, and emoticons facilitate sentiment analysis instead of interfering.

The evaluation in this thesis uses corpora with reviews from Amazon.com, TripAdvisor.com and IMDb.com. The categorization into positive and negative for the Amazon and TripAdvisor follows the scheme suggested by Liu (2012):

$$polarity = \begin{cases} positive, & \text{if rating} > 3 \\ negative, & \text{if rating} < 3 \end{cases} \quad (5.9)$$

In this formula, "rating" refers to the star rating, i.e. the number of stars, as it used by Amazon.com and the circles, as used by TripAdvisor.com. The IMDb corpus already has the labels positive and negative. It is a publicly available corpus<sup>2</sup> created by Pang and Lee (2004).

## Amazon

Amazon<sup>3</sup> is one of the world's leading online retailers. Founded in 1994 as a web-based book shop, Amazon now offers a wide variety of different products. In addition to books, it also sells DVDs, computers and computer supplies, a plethora of electronic devices, but also articles such as pet food or gardening tools. The company also provides means for feedback on products. Buyers of a product can write reviews about products, giving them the chance to recommend a purchase or to discourage from doing so. The review authors express their opinions in continuous text and can summarize their opinion in a star rating (see Figure 5.2). This rating consists of up to five stars. A rating of only one star represents a highly negative opinion towards the product, whereas five stars represent the best grade. Exceptions exist, e.g. by mistakenly choosing

<sup>2</sup><http://www.cs.cornell.edu/people/pabo/movie-review-data/>, last accessed on 24 November 2014.

<sup>3</sup>[www.amazon.com](http://www.amazon.com), last accessed on 24 November 2014.

the opposite rating, or assigning a negative rating for a largely positive review, because one negative aspect severely outweighs all positive aspects. Graduation in between allows the expression of less strong sentiment. A rating of three stars is considered as a neutral opinion towards the purchase, i.e. neither good nor bad (Liu, 2012; Waltinger, 2010).

51 of 59 people found the following review helpful

★★★★★ **Great Tablet...Remember to Read the Specs and Remember What you Pay**, October 12, 2013

By [Badgerx](#) - [See all my reviews](#)

**Verified Purchase** ([What's this?](#))

**This review is from: Kindle Fire HD 7", HD Display, Wi-Fi, 8 GB - Includes Special Offers (Electronics)**

We have 4 kindles in the house now and this one is just as good as the others. For \$139 it has everything one could expect from a tablet. Could it have more memory? Sure, buy the 16G. But since you can download and delete things, just keep what you need on it and it works great.

Help other customers find the most helpful reviews | [Report abuse](#) | [Permalink](#)

Was this review helpful to you?   |

Figure 5.2: A single review on Amazon.

Amazon summarizes all ratings of a product, giving the reader a compact impression on how many people liked/disliked the product (see Figure 5.3). Similarly to products the reviews can be rated themselves. For each review the reader can make a statement about the helpfulness of the review. This mechanism is equivalent to a simple quality management mechanism and guides the reader to the most meaningful reviews. Amazon also contrasts the most helpful positive and negative review, as illustrated in Figure 5.4.



Figure 5.3: Review statistics on Amazon.

The Amazon corpus consists of 2 500 reviews in total, featuring a diverse set of electronic products such as printers. Each round of cross-validation uses 250 reviews for testing and 2 250 reviews for training. The number of positive and negative reviews is balanced, i.e. one half is positive and the other half negative.

## TripAdvisor

TripAdvisor<sup>4</sup> is another valuable source for reviews. On this platform, people can rate their holiday trips. Similar to Amazon's reviews, TripAdvisor has a section for free-text as well as a summarizing rating from 1 to 5 circles (see Figure 5.5). The scale is similar to Amazon's rating scale, i.e. a high number of circles indicates a more positive rating and vice versa.

<sup>4</sup>[www.tripadvisor.com](http://www.tripadvisor.com), last accessed on 24 November 2014.



Figure 5.4: The most helpful positive and negative review of a product.

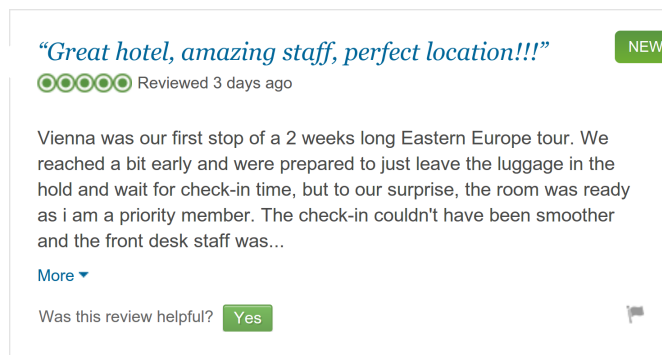


Figure 5.5: A single review on TripAdvisor.

Review statistics are available as well, summarizing the ratings of all reviews of a hotel or point of interest. Furthermore, the summary outlines the ratings regarding certain aspects of a hotel, e.g. the “Sleep Quality”, the “Location”, or the “Service” (see Figure 5.6).

TripAdvisor divides the ratings into fine-grained categories. The middle column of Figure 5.6 shows the classification into “Families”, “Couples”, “Solo”, and “Business”. This allows travellers to easily find reviews related to their traveling type.



Figure 5.6: Review statistics on TripAdvisor.


The TripAdvisor corpus has 1 800 reviews. Each round of cross-validation uses 180 reviews for testing and the remaining 1 620 reviews for training.

## IMDb - the Internet Movie Database

The IMDb corpus has been compiled by Pang and Lee (2004). It contains 2 000 reviews, 1 000 positive and 1 000 negative reviews. For each round of cross-validation, the system uses 100 positive and 100 negative reviews for testing and the remaining reviews for training.

The well-known Internet Movie Database is a platform where users can share their thoughts about movies. Again, a review consists of a textual part as well as a summarizing rating. Here, the rating is more fine-grained on a scale from 0 to 10 (see Figure 5.7). A simple quality management function (“Was the above review useful to you?”) prevents the submission of carelessly written reviews.

6 out of 9 people found the following review useful:



**Central European Western. And a good one!**  
★★★★★

**Author:** bigmac79-1 from Germany  
11 February 2014

This Movie is exceptional German/Austrian Movie. Never saw a Western in this Setting. Sometimes it reminds me of The Great Silence from 1968. But there is something that destroys this incredible Film. The Songs of the Soundtrack are extremely rubbish. What was in Prochaskas Mind, when he choose these Songs? Sad. Great Atmosphere and fine acting, destroyed by a Soundtrack. However, I hope this is the beginning of a new Revival for the European Western. But this time they don't try to pretend that the stories are happening in the US. I can imagine a Western in the Black Forest, or in the flat, wide region of northern Germany, or a polish one.

Was the above review useful to you?

Figure 5.7: A single review on IMDb.

Similar to the other discussed platforms IMDb also has a summary of the ratings. The aggregated rating is easily perceivable in a star, as shown in Figure 5.8.



**Das finstere Tal (2014)**  
Western - 14 February 2014 (Austria)

**Your rating:** ★★★★★★ -/10  
Ratings: 7.7/10 from 517 users  
Reviews: 5 user | 25 critic

Through a hidden path a lone rider reaches a little town high up in the Alpes. Nobody knows where the stranger comes from, nor what he wants there. But everyone knows that they don't want him to stay.

**Director:** Andreas Prochaska  
**Writers:** Martin Ambrosch, Andreas Prochaska, 1 more credit »  
**Stars:** Sam Riley, Tobias Moretti, Helmuth Häusler | See full cast and crew »

Figure 5.8: Review statistics on IMDb.

## 5.5 Results

### TripAdvisor

Applying the contextualization approach to the TripAdvisor dataset resulted in improvements, as compared to the baseline, for the retrieval of precision and f-measure for positive reviews and recall, precision, and f-measure of negative reviews (see Figure 5.9 for detailed results). Four of these results were significant improvements according to Wilcoxon’s rank sum test while the improvement of precision for negative reviews was not significant (see Table 5.1 for the significance of result changes).

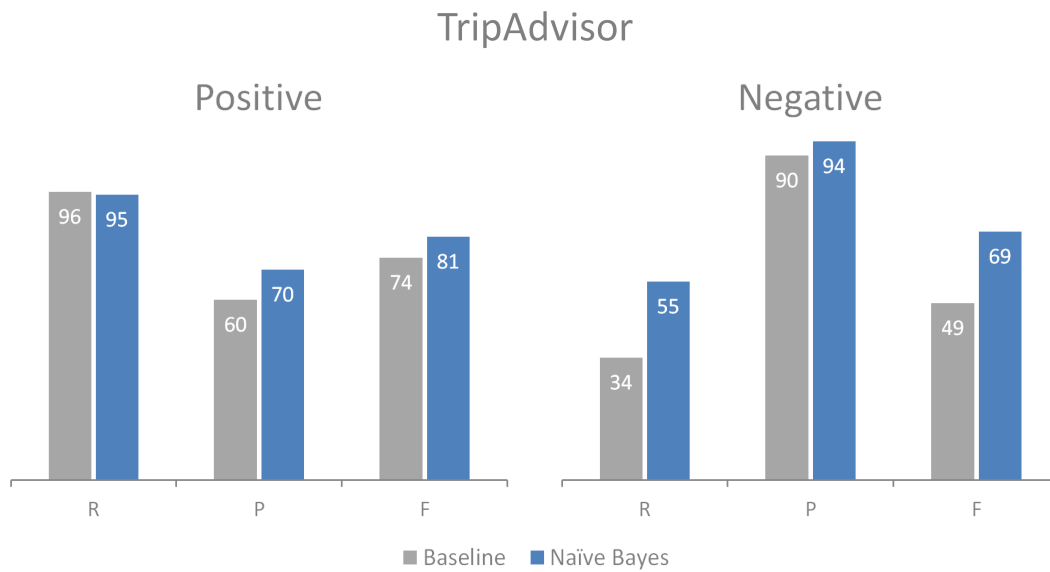


Figure 5.9: Cross-validation on the TripAdvisor data.

The contextualization approach had the strongest gain of performance in recall of negative reviews, by 21 percent points from 34% to 55%, as well as f-measure in negative reviews, by 20 percent points from 49% to 69%. Precision in positive reviews improved considerably as well, by 10 percent points from 60% to 70%.

Recall	Precision	F-Measure
↓	↑	↑
↑	.	↑

Table 5.1: Results for TripAdvisor; an arrow indicates a significant gain/loss, dots indicate stagnation.

There was also one performance loss, in recall of positive reviews by one percent point from 96% to 95%. The performance drop is small but still significant according to Wilcoxon’s rank sum test.

## Amazon

On the Amazon dataset, the baseline performed well in the retrieval of positive reviews and achieved lower results for negative reviews. This trend is in accordance with the baseline results on the TripAdvisor dataset, although on the Amazon dataset the worse performance was less distinct.

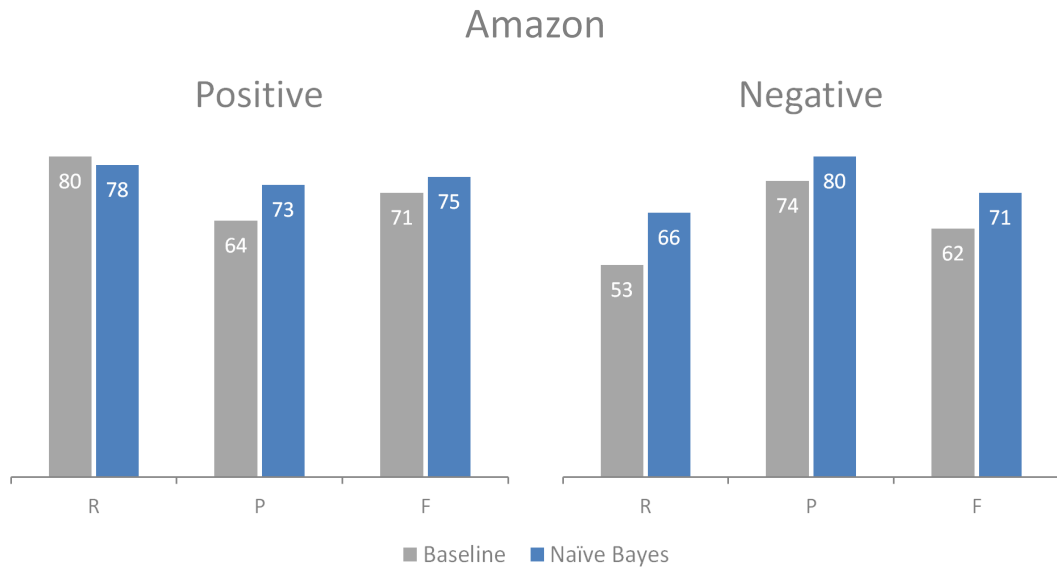


Figure 5.10: Cross-validation on the Amazon data.

The usage of the contextualized lexicon resulted in the improvement of precision and f-measure for positive reviews and recall, precision, and f-measure for negative reviews (see Figure 5.10). The cross-validation showed the highest performance gain in recall for negative reviews, by 13 percent points from 53% to 66%. It showed further remarkable improvements of positive precision by nine points from 64% to 73% and negative f-measure by nine points from 62% to 71%. These three improvements were significant according to Wilcoxon’s rank sum test. For positive f-measure and negative precision the improvement was not significant, i.e. it remains unclear whether the improvement is a matter of chance or can be ascribed to the approach (see Table 5.2 for a summary).

Recall	Precision	F-Measure
↓	↑	·
↑	·	↑

Table 5.2: Results for Amazon; an arrow indicates a significant gain/loss, dots indicate stagnation.



The cross-validation also showed a significant loss for recall when retrieving positive reviews, by two percent points from 80% to 78%. This behavior goes in line with the results observed on the TripAdvisor data.

## IMDb

The contextualization approach performed best on the IMDb dataset. There were five significant improvements in precision and f-measure for positive reviews and recall, precision, and f-measure for negative reviews (see Figure 5.11 for the details). In accordance with the results on the other datasets there was one significant performance loss for recall for the retrieval of positive reviews (see Table 5.3).

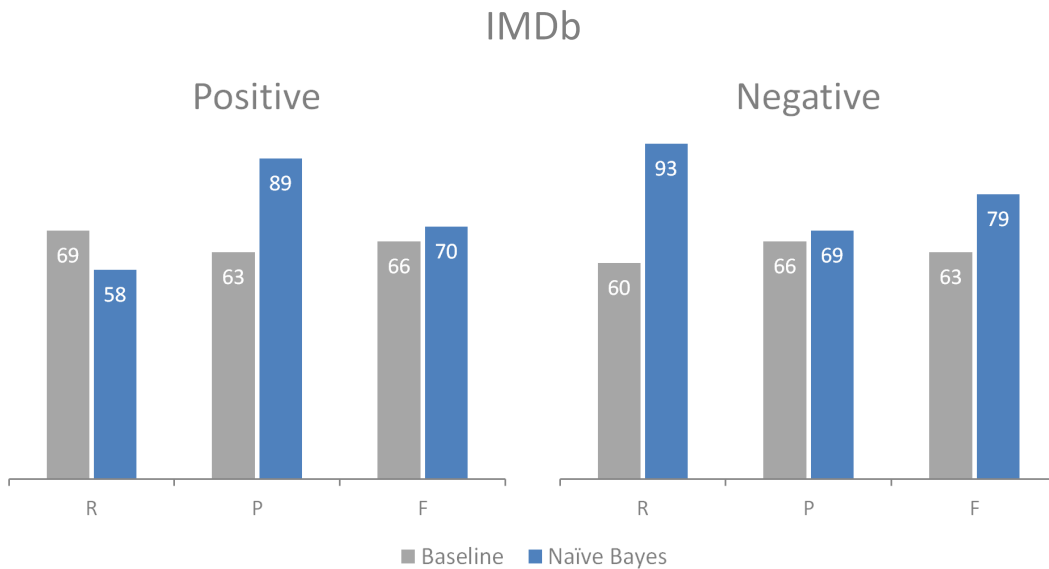


Figure 5.11: Cross-validation on the IMDb data.

The presented approach delivered remarkably superior results for precision in positive reviews and recall and f-measure in negative reviews. The first had an improvement from 63% to 89%, while the second was an improvement of 33 percent points from 60% to 93%. This strong jump resulted in a gain of f-measure of 16 percent points, from 63% to 79%.

Recall	Precision	F-Measure
↓	↑	↑
↑	↑	↑

Table 5.3: Results for IMDb; an arrow indicates a significant gain/loss, dots indicate stagnation.

The performance drop was in recall for positive reviews. The performance dropped by 11 percent points from 69% to 58%, which was also significant according to Wilcoxon’s rank sum test.

## 5.6 Discussion

The achieved results confirm the beneficial effect of context invocation on sentiment analysis. The evaluation showed 15 performance gains in total (see Table 5.4). For these 15 gains Wilcoxon’s rank sum test showed that twelve are significant at the 0.05 level. The non-significant gains were improvements by four percent points in f-measure on positive Amazon reviews, six percent points in precision for negative Amazon reviews, and a four percent improvement in precision for negative TripAdvisor reviews.

Furthermore, the evaluation also showed three performance losses, which were all significant. Interestingly, each of these losses appeared for recall in the retrieval of positive reviews. Two of these losses are minor, i.e. jumps by one and two percent points from 96% to 95% and from 80% to 78%, respectively. The third jump was more considerable. The recall for positive IMDb reviews dropped by eleven percent points from 69% to 58%.

	Total	Significant
Performance gain	15	12
Performance loss	3	3

Table 5.4: Number of total versus significant performance gains and losses.

Table 5.5 contains the summarized percent points of gains/losses of the contextualization method. For instance, the precision of the retrieval of positive Amazon reviews has a value of 9, i.e. there was an improvement of nine points compared to the baseline. Another example is the value of  $-11$  for the recall in positive IMDb reviews. Here, the contextualization method lost eleven points compared to the baseline. The value in the right-most column is the sum of the values in the respective line, while the value below the f-measure is the sum of the values in the respective row.

The contextualization method had three losses in total. Two of them were only minor, i.e. one and two points for recall of positive TripAdvisor and Amazon reviews, while the mentioned loss of eleven points was the worst value.

The contextualization method had the most beneficial impact on recall of negative reviews. Gains of 21, 13, and 33 percent points for TripAdvisor, Amazon and IMDb resulted in a total gain of 67 percent points. The second highest gain was achieved for precision of positive reviews and f-measure of negative reviews. In both cases the gain summed up to 45 percent points.

Interestingly, precision of positive IMDb reviews benefitted most. This stands in strong contrast to the major loss of positive recall for IMDb reviews. The loss of eleven points in recall is compensated with a gain of 26 points for precision, resulting in a gain in f-measure of four points. This gain is statistically significant as well. In summary, despite the major loss in positive

	TripAdvisor	Amazon	IMDb	Total
Positive				
Recall	-1	-2	-11	-14
Precision	10	9	26	45
F-Measure	7	4	4	15
Total	16	11	19	46
Negative				
Recall	21	13	33	67
Precision	4	6	3	13
F-Measure	20	9	16	45
Total	61	39	71	171

Table 5.5: Performance gains and losses per review category and measurement.

recall the contextualization method outperformed the baseline when considering the combined f-measure parameter.

Table 5.6 contains examples for successful contextualization. For instance, the term “busy” is stored with a positive polarity in the sentiment lexicon. However, in connection with “road” the more appropriate polarity is negative. In the sentence “The hotel is located on a **busy road**”, busy is the ambiguous term and *road* is the context term used for disambiguation. Apparently, living in a hotel next to a busy road is undesirable because of excessive noise exposure.

The used sentiment lexicon contains “cool” with a negative value, because the physical experience of cold is unpleasant. However, in colloquial language “cool” has a positive sentiment. The Oxford dictionary defines “cool” as an informal adjective as follows:

“Fashionably attractive or impressive,” e.g.

- “*youngsters are turning to smoking because they think it makes them appear cool,*”

Clearly, in the example sentence in Table 5.6 “cool” also has this meaning and does not refer to a hotel with unpleasant temperature.

Another example is the term “quality”. The Oxford dictionary defines the noun “quality” as follows:

“The standard of something as measured against other things of a similar kind; the degree of excellence of something,” e.g.

- “*an improvement in product quality,*”

“Quality” has a positive value in the used sentiment lexicon. However, in connection with “poor” it conveys a negative opinion, as becomes obvious in the example sentence “Poor quality copies with one edge always dark”.

Discussion remains whether a rule-based approach could also have recognized the polarity switch of “quality”. However, such an approach becomes difficult when the respective terms are at different locations in the sentence. While a simple rule-based approach could easily detect the connection of “poor” and “quality” in the mentioned example sentence a more sophisticated parser becomes necessary in a sentence like “The quality of the printer, despite its good reviews on Amazon, is rather poor”. Another problem is the definition of the rules. To create such rules it is necessary to have a text corpus of sufficient size. A too small corpus bears the risk of missing crucial rules. Creating a rule-based algorithm is also connected with intense manual work. Humans have to read the text data and identify cases where rules might apply. Usually, identifying rules is not a trivial task and cannot be accomplished without a certain skill level in linguistics. Consequently, this task is time-consuming and expensive. Furthermore, rule-based approaches are limited to the stem language of their training corpus. Switching from one language to another involves the compilation of a text corpus in that language. Hiring native speakers or experts in that language becomes necessary to read through the text and create the language-specific rules.

All these reasons make a statistical approach highly desirable. The presented approach does not involve reading through texts and extracting potential rules. Training and testing it on different domains becomes easy and feasible. Moreover, switching to another language merely involves the compilation of a respective corpus before the start of the automatic training procedure.

<b>Ambiguous Term</b>	<b>Sentiment Value</b>	<b>Example sentence</b>
Busy	1	The hotel is located on a <b>busy</b> road.
Complaint	-1	My <i>only</i> <b>complaint</b> would be the service.
Cool	-1	Our room felt like a <i>really</i> <b>cool</b> European apartment with a rooftop terrace.
Expensive	-1	The room was one of the more <b>expensive</b> hotels in Vienna, but still <i>excellent</i> .
Quality	1	<i>Poor</i> <b>quality</b> copies with one edge always dark.
Better	1	Let’s <i>hope</i> they work <b>better</b> .
Cost	-1	Toner <b>cost</b> is way <i>behind</i> competition.

Table 5.6: Examples for successful disambiguation (ambiguous terms are in bold face, context terms in italics).

## Conclusion

*It is paradoxical, yet true, to say, that the more we know, the more ignorant we become in the absolute sense, for it is only through enlightenment that we become conscious of our limitations. Precisely one of the most gratifying results of intellectual evolution is the continuous opening up of new and greater prospects.*

Nikola Tesla

### 6.1 Summary

The main contribution of this work is the creation and evaluation of a software artifact for context-aware sentiment analysis. The approach goes beyond the usage of static sentiment lexicons with mere a-priori polarity values stored for each sentiment term but expands them with dynamically calculated polarity values. The artifact distinguishes between ambiguous and monosemous terms, the latter are treated similar to sentiment terms in static lexicons. For the former, the artifact identifies context terms and stores their probability to occur together with the ambiguous term, in positive or negative texts, respectively. A corpus analysis with a labeled corpus containing positive and negative reviews allows to determine ambiguous terms as well as their co-occurrence frequencies with context terms. The Naïve Bayes technique delivers the mathematical foundation to turn these frequencies into probabilities usable during the application phase.

Applied to a document with unknown overall polarity, the system separates monosemous from ambiguous terms and subsequently determines the context-dependent polarities of the ambiguous terms. After this refinement stage, the polarity values are input to a following sentiment analysis algorithm, in this thesis a keyword lookup algorithm.

A formal 10-fold cross-validation confirms the efficacy of the approach. By involving context knowledge the artifact is able to significantly outperform the baseline algorithm. A set of 2 500 Amazon reviews, 2 000 IMDb reviews, and 1 800 TripAdvisor reviews served as input for the 10-fold cross-validation. Using reviews as evaluation sets is widely known in the literature, despite the limitations of this type of input, e.g. the annotation on the document-level only but not more fine-grained. The high number of statistically significant improvements for recall, precision, and f-measure show the efficacy of the approach.

## 6.2 Discussion

The results achieved with the contextualization approach are promising. The formal evaluation comprised measuring recall, precision, and f-measure on each of the three corpora. Measuring the efficacy for the classification of positive and negative reviews separately gave detailed insight into the performance of the algorithm. This separate evaluation resulted in 18 data points, i.e. recall, precision, and f-measure for positive and negative reviews on three corpora.

The algorithm achieved 15 performance gains in this total of 18 evaluation points. Twelve of these performance gains were significant according to Wilcoxon's rank sum test on the 0.05 level. The most significant improvement was a jump by 33 percent points from 60% to 93% for recall on negative IMDb reviews. The second most significant improvement was a jump by 21 percent points from 34% to 55%, for recall on negative TripAdvisor reviews. The detection of negative reviews benefitted most from the contextualization. Here, the algorithm achieved performance gains for all three parameters on all three evaluation corpora, yielding statistically significant improvements for seven of nine evaluation points.

The 15 performance gains faced three performance losses. The losses occurred only for recall on positive reviews. The most significant performance loss was by eleven points from 69% to 58% for recall in positive IMDb reviews, followed by two points from 80% to 78% in recall for positive Amazon reviews, as well as one point, from 96% to 95%, in recall of positive TripAdvisor reviews.

The total number of performance gains outweighs the number of losses significantly, i.e. 15 gains versus three losses. Given that two of these performance losses are only minor, i.e. a decline by two points and one point, respectively, further confirms the efficacy of the approach.

## 6.3 Future Work

Based on the work accomplished in this thesis there are several paths for future work. The approach, as it is described in this thesis, works on the document level. While this strategy works well on reviews, i.e. short documents, it might fail when document length increases. Topic changes within a lengthy document are a potential source for failure. Going down to the paragraph level, sentence level, or multiple-sentence level might further help to improve the efficacy of the approach. A potential and intuitive strategy is to only consider context terms within a three sentence or five sentence window around the ambiguous term. In other words, only context terms in the preceding or succeeding sentence or the two preceding/succeeding sentences, respectively, will be considered for contextualization.

Another interesting line of research is the application of other machine learners than Naïve Bayes. Support vector machines or linear regression are known to work well in text classification tasks and might deliver results superior to the Naïve Bayes baseline. Alternatively, deep-learning, which currently attracts a lot of research interest, is another future path.

Besides the application of different machine learning algorithms or adding further layers of granularity, integrating already existing resources, e.g. a semantic network such as ConceptNet (Speer and Havasi, 2013), can add further knowledge that has been missed out during the training procedure.

The evaluation in this thesis used three distinct corpora with 2 500, 2 000, and 1 800 reviews per corpus. Given that the approach attempts to learn real-world knowledge this input size is rather small. Expanding the corpus to 100 000 or several millions of reviews will help to iron out bias created by a small corpus size and give valuable insight into ambiguities and context characteristics in large corpora. A further direction is the creation of a corpus with mixed domains. This will give insight on whether there are domain-independent context terms that might be helpful to classify documents in an unknown domain. In this work, the domains are strictly separated into holiday, movie, and product reviews, which results in the problem that machine learning algorithms have in general, i.e. that they are only effective in the domain they have been trained on. Creating a more generic corpus might help to unveil and use features that allow the application of the contextualization in unknown domains without significantly lowering the efficacy of the algorithm.

Sentiment analysis remains a challenging research area. Although significant research effort has been taken it is far from being solved. With the advent of more powerful hardware it will become possible to employ sophisticated techniques such as dependency parsing and machine learning on a large scale and with more complex rule bases and feature sets. This will further raise the efficacy of automatized approaches and push them closer to what humans can achieve. An understanding that comes close to human capabilities requires extensive knowledge. On the one hand, language itself is highly complicated. Complex grammatical structures impose challenges on dependency parsers. Dialects or unclear language further intensify this problem. On the other hand, computers generally lack the extensive world knowledge a human possesses through years of experience and interaction with other humans. Integrating this knowledge into a sentiment analysis toolkit will highly improve its efficacy, but is yet far from being accomplished.

Furthermore, the correct handling of stylistic means such as irony, sarcasm, or metaphors still remains a challenge. Without any further information such as changes in the voice or facial expression, or an extensive knowledge about the statement's topic, these features are hard to detect, even for humans. Adding emoticons or respective hashtags in micro-blogs helps humans to understand them and also proves beneficial for sentiment analysis systems. Without these hints, irony and sarcasm are extremely difficult to understand for computers. Algorithms with this capability will eventually push the machine's understanding close to human skills.





# Bibliography

- A. Agarwal, F. Biadys, and K. R. McKeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 24–32, Athens, Greece, 2009. Association for Computational Linguistics.
- S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC '10, Valletta, Malta, 2010. European Language Resources Association (ELRA).
- B. Baldwin. Cogniac: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ANARESOLUTION '97, pages 38–45, Madrid, Spain, 1997. Association for Computational Linguistics.
- M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.
- P. Beineke, T. Hastie, and S. Vaithyanathan. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Barcelona, Spain, 2004. Association for Computational Linguistics.
- D. Bollegala, D. Weir, and J. Carroll. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1719–1731, 2013.
- E. Brill. A report of recent progress in transformation-based error-driven learning. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 256–261, Plainsboro, New Jersey, USA, 1994. Association for Computational Linguistics.
- A. Burchardt, K. Erk, A. Frank, A. Kowalski, and S. Pado. SALTO - a versatile multi-level annotation tool. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, LREC '06, pages 517–520, Genoa, Italy, 2006.

- E. Cambria, A. Livingstone, and A. Hussain. The hourglass of emotions. In *Cognitive Behavioural Systems*, volume 7403 of *Lecture Notes in Computer Science*, pages 144–157. Springer, Berlin/Heidelberg, 2012.
- E. Cambria, D. Olsher, and D. Rajagopal. Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Conference on Artificial Intelligence, AAI '14*, pages 1515–1521, Quebec City, Quebec, Canada, 2014.
- E. Charniak. Statistical techniques for natural language parsing. *AI Magazine*, 18(4):33–43, 1997.
- E. Charniak and M. Elsner. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09*, pages 148–156, Athens, Greece, 2009. Association for Computational Linguistics.
- K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics, ACL '89*, pages 76–83, Vancouver, British Columbia, Canada, 1989. Association for Computational Linguistics.
- S. Clematide and M. Klenner. Evaluation and extension of a polarity lexicon for german. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '10*, pages 7–13, Lisbon, Portugal, 2010.
- S. Clematide, S. Gindl, M. Klenner, S. Petrakis, R. Remus, J. Ruppenhofer, U. Waltinger, and M. Wiegand. MLSA - A multi-layered reference corpus for german sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC '12*, Istanbul, Turkey, 2012.
- W. W. Cohen. Learning trees and rules with set-valued features. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1, AAAI'96*, pages 709–716, Portland, Oregon, USA, 1996. AAAI Press.
- R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.
- F. L. Cruz, J. A. Troyano, F. Enríquez, F. J. Ortega, and C. G. Vallejo. ‘Long autonomy or long delay?’ the importance of domain in opinion mining. *Expert Systems with Applications*, 40(8):3174–3184, 2013.
- W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. TiMBL: Tilburg memory based learner, version 5.0 reference guide. Technical report, Induction of Linguistic Knowledge Research Group, Tilburg University, 2001.
- K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web, WWW '03*, pages 519–528, Budapest, Hungary, 2003. ACM.

- K. Denecke. How to assess customer opinions beyond language barriers? In *Proceedings of the Third International Conference on Digital Information Management, ICDIM '08.*, pages 430–435, London, UK, 2008.
- X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM '08*, pages 231–240, Palo Alto, California, USA, 2008. ACM.
- Z. Dong, Q. Dong, and C. Hao. HowNet and its computation of meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, COLING '10*, Beijing, China, 2010. Association for Computational Linguistics.
- T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- P. Ekman. Facial expressions. In *Handbook of Cognition and Emotion*, pages 301–320. John Wiley & Sons, Chichester, West Sussex, UK, 1999.
- A. Esuli and F. Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the fifth International Conference on Language Resources and Evaluation, LREC '10*, Valletta, Malta, 2006.
- C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, USA, 1998.
- J. L. Fleiss. Statistical methods for rates and proportions. In *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Chichester, West Sussex, UK, 1981.
- W. N. Francis and H. Kucera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, Massachusetts, USA, 1982.
- A. Gangemi, V. Presutti, and D. Reforgiato Recupero. Frame-based detection of opinion holders and topics: A model and a tool. *IEEE Computational Intelligence Magazine*, 9(1):20–30, 2014.
- L. Garcia-Moya, H. Anaya-Sanchez, and R. Berlanga-Llavori. Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems*, 28(3):19–27, 2013.
- S. Gindl. Different aggregation strategies for generically contextualized sentiment lexicons. In *Workshop on Dynamic Networks and Knowledge Discovery, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Barcelona, Spain, 2010.
- S. Gindl, A. Weichselbraun, and A. Scharl. Cross-domain contextualization of sentiment lexicons. In *19th European Conference on Artificial Intelligence (ECAI)*, Lisbon, Portugal, 2010.
- S. Gindl, A. Weichselbraun, and A. Scharl. Rule-based opinion target and aspect extraction to acquire affective knowledge. In *WWW Workshop on Multidisciplinary Approaches to Big Social Data Analysis, MABSDA '13*, Rio de Janeiro, Brazil, 2013.

- Z. Hai, K. Chang, J.-J. Kim, and C. Yang. Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):623–634, 2014.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference on the European Chapter of the Association for Computational Linguistics*, EACL '97, pages 174–181, Madrid, Spain, 1997. Association for Computational Linguistics.
- J. Henderson and H. Venkatraman. Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal*, 32(1):472–484, 1993.
- A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.
- M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, Washington, USA, 2004. ACM.
- N. Jakob and I. Gurevych. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1035–1045, Cambridge, Massachusetts, USA, 2010a. Association for Computational Linguistics.
- N. Jakob and I. Gurevych. Using anaphora resolution to improve opinion target identification in movie reviews. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 263–268, Uppsala, Sweden, 2010b. Association for Computational Linguistics.
- L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 151–160, Portland, Oregon, USA, 2011. Association for Computational Linguistics.
- N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of First ACM International Conference on Web Search and Data Mining*, WSDM '08, Stanford, California, USA, 2008.
- T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.
- S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

- S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 423–430, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- R. Klinger and P. Cimiano. The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC '14, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.
- U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic. Synesketch: An open source library for sentence-based emotion recognition. *IEEE Transactions on Affective Computing*, 4(3): 312–325, 2013.
- L. Kunze, A. Haidu, and M. Beetz. Acquiring task models for imitation learning through games with a purpose. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS '13, pages 102–107, Tokyo, Japan, 2013.
- T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240, 1997.
- R. Lau, C. Lai, and Y. Li. Leveraging the web context for context-sensitive opinion mining. In *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology*, ICCSIT '09, pages 467–471, Beijing, China, Aug 2009.
- J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morse, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2), 2014.
- W. Lezius. TIGERSearch – Ein Suchwerkzeug für Baubanken. In S. Busemann, editor, *Proceedings of the 6th Conference on Natural Language Processing*, KONVENS '02, pages 107–114, Saarbrücken, Germany, 2002.
- B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
- B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, Chiba, Japan, 2005. ACM.
- B. Lu and B. K. Tsou. Cityu-dac: Disambiguating sentiment-ambiguous adjectives within context. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 292–295, Uppsala, Sweden, 2010. Association for Computational Linguistics.
- N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392, 2013.

- C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, USA, 1999.
- C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, United Kingdom, 2009.
- M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.
- M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
- W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.
- R. Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 869–875, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics.
- S. M. Mohammad and P. D. Turney. NRC emotion lexicon. Technical report, National Research Council Canada, 2013.
- S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 341–349, Edmonton, Alberta, Canada, 2002. ACM.
- T. Nakagawa, K. Inui, and S. Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 786–794, Los Angeles, California, 2010. Association for Computational Linguistics.
- T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, K-CAP '03, pages 70–77, Sanibel Island, Florida, USA, 2003. ACM.
- A. Neviarouskaya and M. Aono. Sentiment word relations with affect, judgment, and appreciation. *IEEE Transactions on Affective Computing*, 4(4):425–438, 2013.
- A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science*, pages 218–229. Springer, Berlin/Heidelberg, 2007.
- A. Neviarouskaya, H. Prendinger, and M. Ishizuka. SentiFul: Generating a reliable lexicon for sentiment analysis. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction*, ACII '09, pages 1–6, Amsterdam, Netherlands, 2009.

- A. Neviarouskaya, H. Prendinger, and M. Ishizuka. SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2(1):22–36, 2011.
- C. Nicholls and F. Song. Improving sentiment analysis with part-of-speech weighting. In *Proceedings of the International Conference on Machine Learning and Cybernetics, ICMLC '09*, pages 1592–1597, Baoding, Hebei, China, 2009.
- J.-Y. Nie, G. Cao, and J. Bai. Inferential language models for information retrieval. *ACM Transactions on Asian Language Information Processing*, 5(4):296–322, 2006.
- G. Paltoglou, M. Theunis, A. Kappas, and M. Thelwall. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 4(1):106–115, 2013.
- B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Barcelona, Spain, 2004. Association for Computational Linguistics.
- B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 115–124, Ann Arbor, Michigan, USA, 2005. Association for Computational Linguistics.
- B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, EMNLP '02*, pages 79–86, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.
- S. Petrov and D. Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, USA, 2007.
- R. Plutchik. The nature of emotions. *American Scientist*, 89(4):344–350, 2001.
- L. Polanyi and A. Zaenen. Contextual valence shifters. In J. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*. Springer, Netherlands, 2006.
- A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 339–346, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.
- S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay. Enhanced Sentinet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.

- G. Qiu, B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27, 2011.
- W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pages 193–197, Torino, Italy, 2009. Association for Computing Machinery.
- A. N. Rafferty and C. D. Manning. Parsing three german treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German, PaGe '08*, pages 40–46, Columbus, Ohio, USA, 2008. Association for Computational Linguistics.
- R. Remus. Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In *Proceedings of the 12th IEEE International Conference on Data Mining, Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE), ICDMW '12*, pages 717–723, Brussels, Belgium, 2012.
- R. Remus, U. Quasthoff, and G. Heyer. SentiWS – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation, LREC '10*, pages 1168–1171, Valletta, Malta, 2010.
- F. Ren and Y. Wu. Predicting user-topic opinions in twitter with social and topical context. *IEEE Transactions on Affective Computing*, 4(4):412–424, 2013.
- E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP '03*, pages 105–112, Sapporo, Japan, 2003. Association for Computational Linguistics.
- J. Ruppenhofer, S. Somasundaran, and J. Wiebe. Finding the sources and targets of subjective expressions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC '08*, Marrakech, Morocco, 2008.
- J. Ruppenhofer, J. M. Struß, J. Sonntag, and S. Gindl. IGGSA-STEPS: Shared task on source and target extraction from political speeches. *Journal for Language Technology and Computational Linguistics*, 29(1):33–46, 2014.
- A. B. Sayeed, J. Boyd-Graber, B. Rusk, and A. Weinberg. Grammatical structures for word-level sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 667–676, Montreal, Canada, 2012. Association for Computational Linguistics.
- R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168, 2000.
- A. Scharl, M. Sabou, S. Gindl, W. Rafelsberger, and A. Weichselbraun. Leveraging the wisdom of the crowds for the acquisition of multilingual language resources. In *Proceedings of the*



- Eight International Conference on Language Resources and Evaluation, LREC '12, Istanbul, Turkey, 2012.*
- H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.
- S. Seebauer. Measuring climate change knowledge in a social media game with a purpose. In *Proceedings of the 5th International Conference on Games and Virtual Worlds for Serious Applications, VS-GAMES '13*, pages 1–8, Bournemouth, UK, 2013.
- K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. *IEEE Intelligent Systems*, 23(3):50–60, 2008.
- R. Speer and C. Havasi. ConceptNet 5: A large semantic network for relational knowledge. In I. Gurevych and J. Kim, editors, *The People's Web Meets NLP, Theory and Applications of Natural Language Processing*, pages 161–176. Springer, Berlin/Heidelberg, 2013.
- P. J. Stone. *The General Inquirer: A Computer Approach to Content Analysis*. M.I.T. Press, Cambridge, Massachusetts, U.S.A., 1966.
- C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC '04, Lisbon, Portugal, 2004*.
- V. S. Subrahmanian and D. Reforgiato. AVA: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems*, 23(4):43–50, 2008.
- J. Surowiecki. *The Wisdom of Crowds*. Anchor Books, New York, USA, 2005.
- M. Thelwall and K. Buckley. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(4):1608–1617, 2013.
- T. Trilla and F. Alias. Sentence-based sentiment analysis for expressive text-to-speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):223–233, 2013.
- P. D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Machine Learning: ECML 2001*, volume 2167 of *Lecture Notes in Computer Science*, pages 491–502. Springer, Berlin/Heidelberg, 2001.
- P. D. Turney. 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, ACL '02*, pages 417–424, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.
- P. D. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical report, National Research Council, Canada, Institute for Information Technology, 2002.

- P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346, 2003.
- K. Veselovská, J. j. Hajič, and J. Šindlerová. Subjectivity lexicon for czech: Implementation and improvements. *Journal for Language Technology and Computational Linguistics*, 29(1):47–61, 2014.
- A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.
- L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.
- U. Waltinger. GermanPolarityClues: A lexical resource for german sentiment analysis. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC '10*, Valletta, Malta, 2010. European Language Resources Association (ELRA).
- C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang. SentiView: Sentiment analysis and visualization for internet popular topics. *IEEE Transactions on Human-Machine Systems*, 43(6):620–630, 2013a.
- D. Wang, S. Zhu, and T. Li. SumView: A web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications*, 40(1):27–33, 2013b.
- W. Wang, H. Xu, and W. Wan. Implicit feature identification via hybrid association rule mining. *Expert Systems with Applications*, 40(9):3518–3531, 2013c.
- X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, pages 1031–1040, Glasgow, UK, 2011. ACM.
- A. Weichselbraun, S. Gindl, and A. Scharl. A context-dependent supervised learning approach to sentiment detection in large textual databases. *Journal of Information and Data Management*, 1(3):329–342, 2010.
- A. Weichselbraun, S. Gindl, and A. Scharl. Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, pages 1053–1060, Glasgow, UK, 2011. ACM.
- A. Weichselbraun, S. Gindl, and A. Scharl. Extracting and grounding contextualized sentiment lexicons. *IEEE Intelligent Systems*, 28(2):39–46, 2013.
- A. Weichselbraun, S. Gindl, and A. Scharl. Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-Based Systems*, 69:78–85, 2014.
- C. Whissell. The dictionary of affect in language. In *Emotion: Theory, Research, and Experience*, volume 4, pages 113–131. Academic Press, Waltham, Massachusetts, USA, 1989.

- J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.
- J. M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.
- F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.
- T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.
- T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.
- M. Wollmer, F. Wengler, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.
- W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probbase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 481–492, Scottsdale, Arizona, USA, 2012. ACM.
- Y. Wu and M. Wen. Disambiguating dynamic sentiment ambiguous adjectives. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1191–1199, Beijing, China, 2010. Association for Computational Linguistics.
- R. Xia, C. Zong, X. Hu, and E. Cambria. Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *IEEE Intelligent Systems*, 28(3):10–18, 2013.
- R. Xu, J. Xu, and C. Kit. HITSZ-CITYU: Combine collocation, context words and neighboring sentence sentiment in sentiment adjectives disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 448–451, Uppsala, Sweden, 2010. Association for Computational Linguistics.
- X. Xueke, C. Xueqi, T. Songbo, L. Yue, and S. Huawei. Aspect-level opinion mining of online customer reviews. *China Communications*, 10(3):25–41, 2013.
- S.-C. Yang and M.-J. Liu. YSC-DSAA: An approach to disambiguate sentiment ambiguous adjectives based on SAAOL. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 440–443, Uppsala, Sweden, 2010. Association for Computational Linguistics.
- J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM '03, pages 427–434, Melbourne, Florida, USA, 2003.

- H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '03*, pages 129–136, Sapporo, Japan, 2003. Association for Computational Linguistics.
- C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01*, pages 403–410, Atlanta, Georgia, USA, 2001. ACM.
- W. Zhang, L. Jia, C. Yu, and W. Meng. Improve the effectiveness of the opinion retrieval and opinion polarity classification. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 1415–1416, Napa Valley, California, USA, 2008. ACM.
- J. Zhu, H. Wang, M. Zhu, B. Tsou, and M. Ma. Aspect-based opinion polling from customer reviews. *IEEE Transactions on Affective Computing*, 2(1):37–49, 2011.

# Curriculum Vitae

## Personal Details

Date of Birth February 26, 1981  
Place of Birth Vienna, Austria  
Nationality Austria  
Title MSc/Dipl.-Ing.  
Address Weissenwolffgasse 12  
1210 Vienna, AUSTRIA



## Education and Career

Diploma Master of Science, 2008, Vienna University of Technology  
Employment Researcher at the Department of New Media Technology,  
MODUL University Vienna, 2008

## Research Areas

Sentiment Analysis, Artificial Intelligence, Natural Language Processing

## Teaching

MODUL University Web Information Systems Development  
MODUL University Tourism and Hospitality Business Applications  
TU Vienna Introduction to Informatics

## Activities

Sabbatical leave with the School of Computer Science and Software Engineering at the University of Western Australia, Perth, in February 2012.

Sabbatical leave with the Natural Language Processing Group, Department of Computer Science, at the University of Sheffield, United Kingdom, in June 2013.

Former co-chair of the GSCL Interest Group on German Sentiment Analysis.

Co-organizer of the 1st Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis (PATHOS 2012), collocated with KONVENS 2012.

Co-organizer of the 2nd Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis (PATHOS 2013), collocated with GSCL 2013.

## Technical Skills

Python, Java, C, Perl

Database design and maintenance, PostgreSQL

Linux Server Administration

Python NLTK, GATE, WEKA, scikit-learn, Flask

Software versioning (Git, SVN, Mercurial)

Extensive knowledge of Microsoft Office

## Languages

German	native
English	academic level
French	basic