SENTENCE-LEVEL SENTIMENT TAGGING ACROSS DIFFERENT DOMAINS AND GENRES

ALINA ANDREEVSKAIA

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

APRIL 2009

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-63357-1
Our file Notre référence
ISBN: 978-0-494-63357-1

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Mrs. Alina Andreevskaia**
Entitled: **SENTENCE-LEVEL SENTIMENT TAGGING ACROSS DIFFERENT DOMAINS AND GENRES**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

______________________________ Chair
______________________________ External Examiner
______________________________ Examiner
______________________________ Examiner
______________________________ Examiner
______________________________ Examiner
______________________________ Supervisor

Approved ______________________________
Chair of Department or Graduate Program Director

__________ 20 ____ ______________________________

Dr. Robin A.L. Drew, Dean
Faculty of Engineering and Computer Science

Abstract

SENTENCE-LEVEL SENTIMENT TAGGING ACROSS DIFFERENT DOMAINS AND GENRES

Alina Andreevskaia, Ph.D.
Concordia University, 2009

The demand for information about sentiment expressed in texts has stimulated a growing interest into automatic sentiment analysis in Natural Language Processing (NLP). This dissertation is motivated by an unmet need for high-performance domain-independent sentiment taggers and by pressing theoretical questions in NLP, where the exploration of limitations of specific approaches, as well as synergies between them, remain practically unaddressed.

This study focuses on sentiment tagging at the sentence level and covers four genres: news, blogs, movie reviews, and product reviews. It draws comparisons between sentiment annotation at different linguistic levels (words, sentences, and texts) and highlights the key differences between supervised machine learning methods that rely on annotated corpora (corpus-based, CBA) and lexicon-based approaches (LBA) to sentiment tagging.

Exploring the performance of supervised corpus-based approach to sentiment tagging, this study highlights the strong domain-dependence of the CBA. I present the development of LBA approaches based on general lexicons, such as WordNet, as a potential solution to the domain portability problem.

A system for sentiment marker extraction from WordNet's relations and glosses is developed and used to acquire lists for a lexicon-based system for sentiment annotation at the sentence and text levels. It demonstrates that LBA's performance across domains is more stable than that of CBA. Finally, the study proposes an integration of LBA and CBA in an ensemble of classifiers using a precision-based voting technique that allows the ensemble system to incorporate the best features of both CBA and LBA. This combined approach outperforms both base learners and provides a promising solution to the domain-adaptation problem.

The study contributes to NLP (1) by developing algorithms for automatic acquisition of sentiment-laden words from dictionary definitions; (2) by conducting a systematic study of approaches to sentiment classification and of factors affecting their performance; (3) by refining the lexicon-based approach by introducing valence shifter handling and parse tree information; and (4) by development of the combined, CBA/LBA approach that brings together the strengths of the two approaches and allows domain-adaptation with limited amounts of labeled training data.

Acknowledgments

This work concludes a long, demanding, yet extremely exciting journey that started with my studies in Linguistics, continued in Computer Science, and finally brought these two domains together in my current research in Natural Language Processing (NLP). Studying at Concordia and being part of Montreal's research community was undoubtedly instrumental in the shaping of my research interests and understanding of NLP.

The completion of this thesis would not have been possible without the assistance of my supervisor, Professor Sabine Bergler. I am very thankful to her for the invaluable help and support she provided. I particularly appreciated our long casual discussions. Despite her enormous workload and constraints on her time, she always made herself available to me and other students and worked extensively with me on the revisions of this thesis and of papers that it is composed of. Her dedication, encouragement, and patience allowed me to develop and try new ideas. Working with her was a great learning experience.

I would like to express my deep appreciation to the members of my doctoral committee, Professors Leila Kosseim, Thiruvengadam Radhakrishnan, Ching Y. Suen, and Charles Reiss for their valuable feedback on my thesis. I am particularly thankful to Professor Kosseim. Her course on statistical methods in NLP had shaped my understanding of approaches to NLP. I am also indebted to her for her thorough comments on my dissertation and for her advice and support in my job search. I am also grateful to Professor Radhakrishnan for his unfailing support and insightful advice throughout my studies.

I would like to express my gratitude to the external examiner, Professor Diana Inkpen for her very thorough comments and for finding time in her busy schedule to read my thesis in a very tight timeline.

I want to say special thanks to my family for their affectionate support throughout my PhD studies and specially to my husband, Alex Bitektine, whose dedication, wisdom and expertise in research methods made him an indispensable member of my support team. His understanding and interest in my work provided me much needed support to face the challenges of the doctoral program. I am grateful to my parents for their encouragement and faith in me.

I am also indebted to Professor Ariadna I. Kuznecova from Moscow State University who supervised my Master's thesis in linguistics and to whom I own the research skills and passion for computational linguistics.

The completion of this work would not have been possible without the financial support from NSERC, Concordia University, and my supervisor.

I would like to thank the administrative staff of the Department of Computer Science and Software Engineering, and especially Ms. Halina Monkiewicz for her friendly help during my graduate studies.

I would like to thank my fellow graduate students, especially the teams of ClaC labs, for their empathy and a great time that we had together.

Finally, I would like to recognize and express my deep appreciation to a number of researchers in a broader academic community whose feedback on the earlier versions of parts of this thesis was indispensable for the development of the manuscript presented here.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traditional Natural Language Processing (NLP) applications mostly concentrate on topical text characterization that deals with the communicated facts and objective presentation of the information. Nevertheless, opinions, sentiments or attitudes expressed in the texts often constitute the main informational content of the text, the main motivating factor for its creation. For many different categories of users, the expressed sentiment is also relevant in a given document. The common examples of such users are shoppers (or computer agents on their behalf) looking for products, politicians and corporations interested in public sentiment and the tone of their coverage in mass media, or researchers in political science and sociology studying public opinion dynamics as reflected in the media.

The importance of sentiment information for the end users should not be underestimated. Surveys [29, 55] cited by Pang and Lee [93] show that between 73% and 87% of readers of online reviews in US have been influenced by the opinions expressed in these reviews and resulting sellers' reputation. Search engines started to include sentiment analysis capabilities in their applications (e.g., the pilot project for Microsoft Live search, that uses the Hu and Liu [56] approach). Special portals have been created to enable the search for "opinionated" [144] blogs [18], twitters `http://twitrratr.com/`, and product reviews [113]. A number of approaches have been developed for sentiment-based summarization of reviews, for information extraction combined with sentiment analysis, and for multiple-perspective question answering that augments traditional QA with subjectivity analysis. Other systems track political opinions and reactions to events in the press or in blogs (e.g., `http://www.jodange.com/`). In the academic community, researchers in political science, sociology, and management are now using sentiment analysis in order to study the dynamics of public sentiment [8].

A recent surge in interest towards sentiment analysis both in the research community and in the industry was prompted by an increasing demand for information about sentiment and opinions expressed in texts, as well as by availability of large amounts of on-line data that requires efficient processing. The research in this direction was further stimulated by the availability of large sets of texts, often already tagged with sentiment or rating by their authors, and the development of new, more efficient machine learning techniques [93]. These factors have led to a paradigmatic shift in the research on sentiment and subjectivity from mostly linguistic approaches motivated by the studies on perspectives in narrative (e.g., [138]) and small pilot studies [117, 52, 118] to application of machine learning techniques that can achieve high precision when large amounts of training data are available. In the recent few years, a considerable amount of work has been done in sentiment analysis of such domains as movie and product reviews, where it was relatively easy to obtain large quantities of labeled data. This research produced highly accurate (up to 90%) in-domain trained systems. However, as researchers started to extend sentiment tagging to other domains or to work on more general applications, there came a realization of the shortcomings of crude machine learning approaches in real-life contexts, where it was no longer possible to rely on the availability of large training corpora from the same domain and genre. Domain adaptation and system portability, thus, have recently emerged as the most salient challenges in sentiment analysis [46, 102].

The text in Figure 1 contains several instances where knowledge about positive and negative sentiment expressed by individual words is not sufficient. In Figure 1 words that express positive and negative sentiment are underlined (e.g., crisis, pleased, terrific). Consider processing sentence $< S9 >$:*Although two candle fires were reported, no one was injured and no crime spikes occurred following the blackout, the mayor reported*$< /S9 >$. Automatic sentiment determination systems based on unigrams (acquired from training corpora or lexicons) are likely to be misled by the presence of several negative clues in this sentence and would tag it as negative: fires, injured, crime, and blackout. The standard automated sentiment analysis will not take into account other elements of the text that influence its interpretation by humans — valence shifters [99] like *although* and *no* that reverse the sentiment of the negative markers *fire, injured, crime*, making it an overall positive sentence. This shows that, apart from unigrams (words), which are traditionally used by sentiment analysis applications, a successful sentiment tagging system should also make use of special handling rules for words and expressions that, while not being sentiment-bearing themselves, still can influence the sentiment of the expression (e.g., negations, scalar modifiers, etc.). Different approaches to sentiment tagging deal with this issue of lexical sentiment modifiers in different ways: in supervised corpus-based approaches, unigrams can be replaced by higher-order n-grams that are expected to capture some relevant syntactic

$< S1 >$“All power is restored,” a tired but relieved Mayor David Miller announced at 9:40 p.m. as he praised Torontonians’ calm under crisis during a blackout that had enveloped Toronto’s west end. $< /S1 >$ $< S2 >$“My feeling right now is relief,” Miller told media gathered at City Hall last night, almost 24 hours after the power outage was announced. $< /S2 >< S3 >$“It went a little bit faster than we hoped,” he said, adding ”it will take a little bit of time for people’s houses to be warmed up. ” $< /S3 >$

$< S4 >$“I thought Torontonians were terrific. $< /S4 >< S5 >$ They were calm. $< /S5 >< S6 >$ They were helpful. $< /S6 >< S7 >$They helped their neighbours. $< /S7 >< S8 >$They let the city know where there were issues,” he said. $< /S8 >$

$< S9 >$Although two candle fires were reported, no one was injured and no crime spikes occurred following the blackout, the mayor reported. $< /S9 >$

$< S10 >$Although the mayor was pleased the heat was on, he cautioned residents about turning on their power in stages, “cautiously and gradually” giving houses a chance to warm up gradually and the power system a chance to stabilize “to minimize the load that will occur all at once on the power system ... so there are no further blackouts or brownouts. ” $< /S10 >$

Figure 1: Fragment of newspaper text with subjective elements. For easier reference each sentence is assigned a number shown in angle brackets.

patterns, while in lexicon-based systems, valence shifters and information about syntactic dependencies can be included into the set of features, used by the system.

1.1 Sentiment Annotation of Different Genres

The present study explores sentence-level sentiment annotation of four very different genres: movie reviews, product reviews, blogs, and news. Of these four genres, movie reviews, product reviews, and, to lesser extent, blogs have been in the focus of text-level sentiment analysis for several years. At the same time, news remain relatively unexplored both at text and sentence levels.

Each domain and genre poses specific challenges for sentiment annotation. Movie reviews often combine neutral sentences that describe the film plot and the sentiment-laden ones. Product reviews are very domain-specific, and systems trained on one domain (e.g., kitchen appliances reviews) will not perform well on some other domain (e.g., reviews of DVD’s) [16].

VICTORY AND DEFEAT

Metronews, Vancouver, November 14, 2008 $< S1 >$*Mounting public anger over the recently passed gay marriage ban in California turned toward the Mormon church, which poured millions of dollars into advocating the passage of Proposition 8.*$< /S1 >$ $< S2 >$*However, gay marriage advocates in Connecticut had reason to celebrate after a judge cleared the way for homosexual marriages to begin, prompting gay couples to immediately march to New Haven City Hall to tie the knot.*$< /S2 >$

Figure 2: Example of mixed-sentiment newspaper text. For reference each sentence is assigned a number shown in angle brackets.

Blogs differ from other genres by their highly emotional context, colloquial style, a careless presentation (typos, agrammatical sentences), as well as the use of emoticons. They are sometimes annotated with author's mood, but these mood labels are very diverse (there are several hundreds of them) and are used inconsistently.

Newspaper texts present a particular challenge for sentiment analysis: newspaper articles usually present a "balanced" view on their topic, combining different, often conflicting, opinions, citing opponents, and presenting both objective facts and subjective "points of view", or, as in the Figure 2, describe different news events in one text. The text in Figure 2 also demonstrate that there is a fine boundary between objective and subjective material in newspaper texts. Many facts, even presented in an objective way, elicit an emotional response from the reader — "good" vs. "bad" news. Many texts also present opinions and sentiments that do not belong to the article's author [93]. It has become standard to consider such examples as subjective, following the definition given in [145], where all sentences that contain subjective elements (e.g., words and phrases) are deemed subjective. Another approach would be to combine the sentiment annotation with detection of opinion holders (sources) [63]. At the present time, however, the opinion holder detection is a separate task that requires different methods and those results can considerably influence the outcome of subsequent sentiment annotation. Therefore, opinion holder detection is out of scope of this dissertation.

The heterogeneity of sentiment often found in different parts of a single newspaper text dictates the need for splitting newspaper texts into smaller parts (sub-document units) with common sentiment. Therefore, in this dissertation, sentiment analysis will be performed **at the sentence level** rather than at the level of the entire text. Sentiment tagging of small language units avoids the problem of sorting out different opinions that may co-exist in a text, but at the same time, it poses other challenges that make it more difficult than sentiment analysis of sentimentally homogenous texts. The relatively small size of sentences means that the decision about the sentiment of a sentence has to be made based on a small number of sentiment clues and, thus, is more sensitive to system errors and to model sparseness.

The development of sentence-level annotation as a subdomain in sentiment research opens up an opportunity for development of a number of applications in text mining and information retrieval. First, the identification of sentiment expressed by opinion holders in relation to a certain topic, event or issue cannot be reduced to the sentiment of the whole text and requires that fine-grained level of annotation that sentence-level research offers. Second, practical applications of sentence-level annotation also include the study of hedging in scientific literature, information retrieval applications that often require sentence-level processing, summarization, and text categorization.

The characteristics of newspaper texts described above pose particular challenges for sentiment annotation systems. First, corpus-based supervised machine learning approaches that have been successful on other genres of texts are less reliable on newspaper data. While these methods have reached high accuracy on movie and product reviews, where large amounts of annotated data are easily available for training, the performance of supervised machine learning approaches on newspaper texts is compromised by scarcity of annotated newspaper text corpora. Sentiment of newspaper texts used for training has to be assigned manually, which requires substantial annotator efforts and usually yields relatively small datasets for training and testing of supervised approaches. Moreover, newspaper texts are very diverse in their topics, which range from politics and finance to sports and art. Therefore training a classifier for all these sub-domains is a much more difficult task than classifier training for more homogenous movie or product reviews corpora. Thus, supervised corpus-based machine learning approaches, which are a method of choice for sentiment analysis have only limited applicability in detection of sentiment of newspaper texts. This dissertation seeks to address the challenges posed by sentiment annotation of newspaper texts through the development and testing of novel approaches to this task that would combine the strengths of corpus-bases statistical methods with portability and robustness of the lexicon-based approach.

1.2 Thesis Motivation

The motivation for this research stems both from the unmet needs in applied research focused on creation of high-performance sentiment taggers for commercial and research applications, and from pressing theoretical questions in NLP, where the exploration of limitations, synergies and the potential of domain-dependent supervised corpus-based machine learning approaches and more general lexicon-based methods in sentiment annotation remains practically unaddressed.

On the **applied research side**, several areas of research in NLP can benefit from a reliable approach to classifying texts and text spans into those expressing positive, negative and neutral/mixed sentiment:

- Multiple Perspective Question Answering (MPQA) [140] that aims at finding a range of opinions being expressed in the world press about a given topic and clustering opinions and their sources;
- Summarization [53];
- Sentiment tagging in order to identify positive and negative product reviews [30, 56, 100, 160] and to generate a list of product attributes and aggregate sentiment about them [32];
- Tracking public sentiment about political events [35];
- Identifying support and opposition in parliamentary debates [129];
- Recognizing flames and hate messages [147, 130, 117];
- Clustering messages by ideological point of view [109];
- E-mail processing [76].

Outside the area of NLP, user interfaces, recommender systems and software agents would benefit from non-topical text characterization that would allow the development of agents capable of collecting and summarizing sentiment of users towards products and services (in a consumer report fashion) and providing the users with informed suggestions in respect to product selection, product quality and features. Sentiment analysis can also assist in reputation management and public relations, business and government intelligence, sociology and political science.

Besides multiple practical applications of sentiment analysis for various categories of users, the problem of sentiment tagging has important **theoretical implications** for research in computational linguistics and natural language processing. One of the main challenges faced by natural language processing as a discipline is the trade-off between the richness of linguistic data available through manual annotation and tagging on one hand, and the enormous expense (in terms of time, manpower and other resources) required to produce comprehensive annotations of the entire language system. Because of these constraints, NLP researchers have to consider a trade-off between knowledge-poor corpus-based methods, which rely on fairly "inexpensive" machine learning techniques, and knowledge-rich, but "expensive" approaches that require extensive linguistic analysis. While supervised corpus-based systems can achieve relatively high performance when trained and evaluated on annotated texts from the same domain (e.g., movie reviews [95]), their performance rapidly deteriorates when training data is limited or when a system trained in one domain is used to annotate texts from another domain (e.g., a system trained on movie reviews and used to annotate newspaper texts) [11, 102]. These systems, thus, remain *domain-dependent*. Knowledge-rich, lexicon-based systems, on the other hand, are based on general lexicons, grammar rules, and heuristics that apply equally to all domains and genres, but ignore the specificities of each of the domains. Such systems, therefore, perform worse than supervised corpus-based machine learning system systems with sufficient in-domain training, but better than knowledge-poor systems with insufficient in-domain data or with out-of-domain training and display a fairly stable performance across different domains and genres. Lexicon-based approaches, thus, tend to be *domain-independent*.

The issue of domain-dependence has become a major problem for state-of-the-art sentiment tagging systems. So far, the solution to this problem has been sought within corpus-based learning and lexicon-based paradigms separately. The work presented here seeks to propose a solution that takes advantage of the strengths of the two approaches.

This study, thus, seeks to combine the breadth and comprehensiveness of corpus-based machine learning approaches with the benefits of detailed semantic information extracted using manual and automated data acquisition methods at different linguistic levels.

The study presented in this dissertation starts with systematic analysis of factors that influence the performance of corpus-based statistical approaches to sentiment tagging. This set of experiments seeks to fill the gaps in the current studies of sentiment tagging by conducting an investigation of a large spectrum of factors that influence the performance of supervised machine learning approaches. This step is believed to be critical for the establishment of a valid baseline for the evaluations of the approaches developed in this dissertation. The study continues with the development of a system for sentiment marker extraction from WordNet glosses, followed by the integration of the acquired list of words in a lexicon-based system. The final stage of this research attempts integration of the lexicon-based system with a corpus-based system in an ensemble of classifiers. Figure 3 reflects the

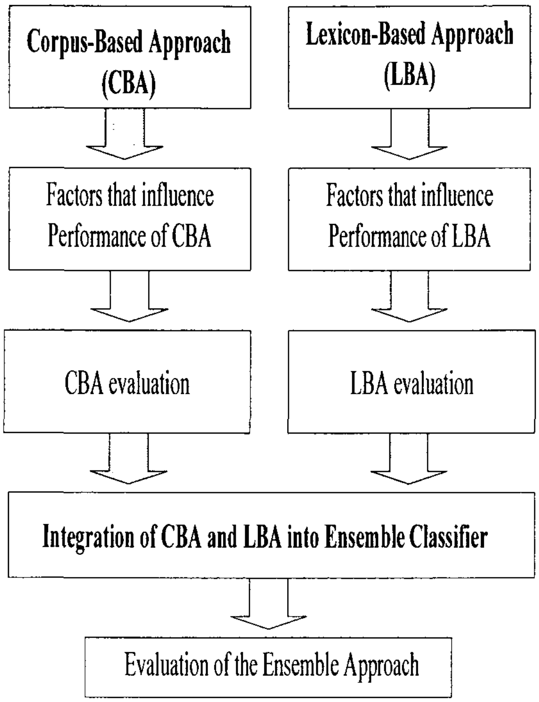
Corpus-Based Approach (CBA)
Lexicon-Based Approach (LBA)
Factors that influence Performance of CBA
Factors that influence Performance of LBA
CBA evaluation
LBA evaluation
Integration of CBA and LBA into Ensemble Classifier
Evaluation of the Ensemble Approach

Figure 3: Thesis structure.

logical flow of this research.

The approach is evaluated on the data from four different domains and genres: news, movie reviews, product reviews, and blogs. Multiple experiments with different system components developed in the course of this research have been presented at several international conferences including EACL-2006, AAAI-2006, Senseval-2007 and ACL-2008 and are reported here in Chapters 4 through 5.

1.3 Evaluation Approach

Throughout this study, the evaluation of different systems is guided by a set of principles, introduced to ensure rigor and consistency of the evaluation procedure:

- **Use of both in-domain and out-of-domain training:** The use of both in-domain and out-of-domain training was motivated by a special attention to the issue of system portability across domains. It has been observed, for example, that corpus-based supervised machine learning systems (CBSs) do not perform well when training data is scarce or when it comes from a different domain [11, 102], topic [102] or time period [102]. Given that these are very common real-world constraints on training data quality and availability, the problem of system portability across different domains becomes a serious issue for practical applications of corpus-based approaches in sentiment annotation.

- **Evaluation of not only the final ensemble system, but also of each of the system components** The individual evaluation of the CBS and LBS allowed the establishment of performance baselines that served as benchmarks in the evaluation of the performance of the combined (ensemble) approach and allowed the assessment of gains associated with this approach and with the precision-based voting technique.

- **Use of multiple domains in the evaluation:** The accuracy of both manual and automatic sentiment annotation has been shown to be highly dependent on the domain/genre of the texts on which the evaluation is conducted (see Chapters 2, 3, 4). As a result, the performance of an approach, for example, on movie reviews cannot be meaningfully compared to the performance of another approach on a different domain. Moreover, given substantial linguistic, stylistic, and structural differences among texts of different domains, a method's good performance on one domain may not necessarily translate into an equally good performance on another domain. Thus, the decision to conduct evaluation on multiple domains provides greater confidence in generalizability of findings and reliability of the reported results.

In the study presented here, sentiment is understood as a ternary category that includes positive, negative, and neutral classes. Therefore the experiments presented here and in the following chapters classify sentences into three classes wherever the data permits it (e.g., news and blogs). Pang and Lee [95] have reported improvement in the accuracy of classification of movie reviews *texts* when objective sentences were first eliminated from the texts and then the remaining subjective sentences were used to classify the texts they belonged to as positive or negative. Since the technique used by Pang and Lee, however, is applicable only at the text level, it was not used in application to sentence-level text annotation in the research presented here.

The performance of the approaches presented here is measured using four common performance measures: accuracy, precision, recall, and F-measure [78]. Depending on the component and dataset, accuracy is computed for binary (positive vs. negative) and ternary (positive vs. negative vs. neutral) classification. Ternary classification accuracy is measured as a percentage of correct labels for all three categories out of the entire size of test set and is computed as:

$$Acc_{ternary} = \frac{correct\_positives + correct\_negatives + correct\_neutrals}{all\_data}$$

Binary classification performance is evaluated by its accuracy, precision, recall, and F-measure. Binary accuracy is computed as the percentage of correctly assigned positive and negative labels over the number of all sentences with positive and negative labels in the test gold standard.

$Acc_{binary} = \frac{correct\_positives+correct\_negatives}{gold\_positives+gold\_negatives}$

The definition of precision and recall extend the definitions[1] of given in Manning and Schütze [78] to the two-category case. Precision of binary, positive/negative classification is defined here as a proportion of correct positive and negative labels given by the system over the number of all positive and negative labels assigned by the system. Sentences that were not tagged as positive or negative are ignored:

$P = \frac{correct\_positives+correct\_negatives}{all\_positives\_assigned+all\_negatives\_assigned}$

Recall is the percentage of correct positive and negative labels assigned by the system over the sum of positives and negatives in the gold standard:

$R = \frac{assigned\_positive+assigned\_negative}{gold\_positives+gold\_negatives}$

Finally, F_1-measure is computed based on precision and recall:

$F_1 = \frac{2PR}{(P+R)}$

Statistical significance of the results was evaluated with Student's t-test, which is a de-facto standard in NLP and is widely used in the comparison of algorithm performance [2].

1.4 Intended Contributions

The specific theoretical and methodological developments proposed in this document will contribute to several NLP research areas:

- Development and testing of algorithms for acquisition of sentiment annotated words based on dictionary definitions that would complement the currently existing methods of automatic lexical acquisition. Given the lexicographic properties of dictionary definitions, this was expected to provide substantial improvements in performance of the word acquisition system;

- systematic study of several approaches to sentiment classification. The analysis, conducted on several genres and domains and at different levels of the language, leads to a deeper understanding of the role of different methods for sentiment classification.

[1] Precision is defined as true_positives/(true_positives+false_positives) and recall = true_positives/(true_positives+false_negatives).

- insights into relative advantages and limitations of supervised machine learning and knowledge-rich approaches to sentiment classification;
- refinement of the lexicon-based approach by introduction of valence shifter handling and parse tree information in sentiment annotation; and
- development of a combined approach, that would bring together the strengths of knowledge-poor (corpus-based machine-learning) and knowledge-rich (lexicon-based) methods of sentence-level sentiment classification using a novel weighting scheme.

The two important theoretical developments in this dissertation — the method for automatic lexical acquisition from dictionary definitions of sentiment-bearing words and an approach that combines unsupervised, lexicon-based and supervised corpus-based learning methods of sentiment tagging — are grounded in linguistic theory. The first contribution, the development of a lexical acquisition method, takes advantage of the properties of dictionary entries as a special kind of structured text that has some important advantages over other types of texts commonly used in lexical acquisition. The second contribution, the development of an approach to sentiment annotation that combines the benefits of corpus-based and lexicon-based methods, builds upon the theorized complementarity of domain-specific and general knowledge in human cognition and provides a promising new direction for research on domain adaptation and system portability.

In addition, the dissertation contributes to Natural Language Processing by exploring the issues of domain adaptation and system portability across different domains, the factors influencing performance of corpus-based and lexicon-based methods in text classification, the role of valence shifters and syntactic information in lexicon-based sentiment annotation, the use of hypernym relations in word sense disambiguation, and finally, developing a precision-based voting method for classifier integration.

The theoretical developments outlined above are supported by empirical work on three sentiment-tagging systems using machine learning, lexicon-based, and combined approaches, respectively. The output from the developed word acquisition module and the sentiment tagger systems are used here to evaluate the effectiveness of the proposed theoretical approaches and methods and to identify the directions for further research in this domain.

1.5 Thesis Structure

The dissertation is organized as follows. The current state-of-the-art sentiment and subjectivity analysis approaches are reviewed in Chapter 2. That chapter also clarifies the terminology used in the non-topical text categorization research. Chapter 3 explores different factors that influence the performance of corpus-based machine-learning approaches and identifies a corpus-based method that has the best performance on newspaper texts. Chapter 4 introduces an unsupervised, dictionary-based approach to sentiment tagging of words and their senses, evaluates its performance within a lexicon-based sentiment tagging system, and explores the role of valence shifters and syntactic information in lexicon-based approaches to sentiment-tagging. Chapter 5 presents a novel approach that combines the two methods of sentiment tagging: a supervised corpus-based machine-learning approach and unsupervised, lexicon-based approach using precision-based voting weighting scheme. Finally, Chapter 6 summarizes the results and delineates the directions for future research.

Chapter 2

Background

2.1 Introduction

Subjectivity and sentiment analysis consists of a "computational treatment of opinion, sentiment, and subjectivity in text" [93]. It seeks to classify language elements at different levels — from words and their senses to texts and groups of documents — according to the opinion, emotion, or sentiment they express. Systems that are able to identify sentiment of a text or its components have a number of applications: multiple-perspective question answering, summarization, information extraction, etc.

The task of sentiment and subjectivity analysis has attracted considerable interest since the 1990s when the first automatic systems for these tasks were developed [117, 148, 149, 22]. Since then it has become a major research stream within NLP. The current work in this area stems from the research in *content analysis* [71] and *point-of-view tracking* in narrative [12, 133, 138]. It is closely related to research in *affective computing* [97][1] and *directionality*, i.e., Is the agent in favor of, neutral, or opposed to the event? [54].

Sentiment research now covers not only English, but a number of other languages: Japanese [69, 59], Chinese [153, 155], and Romanian [120]. In the following, the state of the art of current research on English sentiment tagging/opinion mining will be presented.

Similar to other emerging fields of research, the terminology in sentiment analysis has not stabilized yet and even the very definition of sentiment can be problematic [93]. The terms *sentiment* [31, 94, 62], *semantic orientation* [53, 147, 60, 130, 131, 52], *polarity* [53, 147, 111, 52][2], *opinion* [14, 63], *valence* [99] and *attitude* [9] are used to relate to the same or similar concepts. These terms are often used without any formal definition, either as synonyms

[1] *Affective computing* is defined by Picard [97] as computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.

[2] The last two terms are often used interchangeably and are defined in the same way.

or to refer to different aspects of the phenomenon (for example, [62] define sentiment as *affective part of opinion.*) Some of these terms already have a different meaning in linguistic tradition (e.g., polarity, valence) and therefore are confusing.

Given that the focus of this research is on capturing of sentiment expressed in a text as positive, negative, or neutral (or mixed), I will refer to this domain of research as ***sentiment analysis***. This term is preferred here because (1) it is not associated with any other research tradition and, thus, avoids the potential for confusion with the research in other areas (as it is the case with *polarity tagging*), (2) it accurately reflects the type of information being extracted from the texts (as opposed to, for example, opinions that may have also a topical component), and (3) it is parsimonious and precise.

The task of separation of neutral, objective, non-opinionated sentences and texts from sentiment-laden, subjective ones has been actively explored in a separate, yet closely related field of subjectivity analysis. The *subjectivity analysis* research goes back to work by Banfield [12] and Wiebe [138] and is focused on drawing a distinction between 'subjective' words and texts that present opinions and evaluations, on one hand, and 'objective' words and texts, used to present factual information on the other [147, 145, 143]. *Sentiment analysis*, thus, differs from *subjectivity analysis* by the set of categories into which these two analyses classify language units: subjectivity analysis is concerned with the division into subjective and objective categories, while sentiment analysis aims at dividing them into positive, negative and sometimes neutrals.

There is a considerable overlap between the *objective* category in subjectivity analysis and the category of *neutrals* in sentiment analysis. Similarly, the subjective category in subjectivity analysis to a great extent overlaps with the combined categories of positives and negatives in sentiment research, for instance, the following example of a subjective sentence from [143] will be considered negative in sentiment analysis: *"Western countries were left frustrated and impotent after Robert Mugabe formally declared that he had overwhelmingly won Zimbabwe's presidential election"*. Nevertheless, these categories are not exactly the same: for instance, words that reflect private state and, hence, are subjective, may not necessarily be positive or negatives, e.g., private state words such as *feel, emotion, interest, surprise* are subjective in subjectivity classification, but neutral in sentiment classification. Similarly, positive or negative news can be "good" or "bad" objectively (e.g., factual, objective information about a natural disaster). In this latter case, however, drawing a line between subjective and objective is much more difficult since the objective information may still elicit reader's emotions and the perception of news as good or bad is often defined by reader's views, values, and background (e.g., the same outcome of a sports game can be good news for the fans of the winning team, and bad news for the looser). For the purpose of this dissertation, any text that has an emotional component, intentional or not, is considered subjective[3].

In recent years, the field of sentiment analysis changed its focus from binary (positive-negative) classification [52, 131, 132] to a classification that includes neutrals as a third category [70, 62]. Empirical observations show that separating positives or negatives from neutrals is a much more difficult task than the differentiation of positive elements from negative ones: most of the errors produced by automatic systems and most of the disagreement between human annotators involve separation of neutral words, sentences, or texts from either positive or negative ones [62]. The very definition of "neutrals" also poses a problem[4]. In the literature, "neutrality" has a dual meaning: "lack of opinion" [159], or "a sentiment that lies between positive and negative" [93]. The former definition is mostly used in the area of subjectivity analysis, while sentiment analysis favors the latter interpretation. In this dissertation, the term will be used in this latter meaning.

In addition to sentiment and subjectivity analysis, a number of other tasks related to these major research streams have emerged in the last decade: assignment of more fine-grained affect labels to language elements based on various psychological theories [134, 121], detection of opinion holder [62, 63, 66, 26, 14, 69] and target [57, 47, 56, 100, 67, 69], perspective [75], pros and cons in reviews [65], assignment of ratings to movie reviews [96], identification of support/opposition in congress debates [129], prediction of election results [68], detection of bloggers' mood [87, 86, 73], happiness [84], politeness [108], assessment of review quality [67, 160], and other. The analysis of these subtasks, however, is beyond the scope of this review.

In the following subsections I will review the current research in sentiment and subjectivity analysis of these linguistic units — words, sentences, and texts.

2.2 Word-level Sentiment and Subjectivity Analysis

It has been widely recognized that semantic properties of individual words, such as word sentiment, are good predictors of semantic characteristics of a phrase or a text that contains

[3] Implicitly, a similar view was adopted by the organizers of the SensEval-4 Affective Text task where headlines were often annotated as positive or negative only because they referred to events that elicited an emotional response from the annotators.

[4] In their experiments in manual annotation, Kim and Hovy [62] observed that attempts to draw a distinction between non-opinion and neutral sentiment resulted in confusion for human annotators. In a more detailed study, Koppel and Schler [70] observed that human perception of neutrality depends on the text genre. For instance, human annotators tagged as neutral posts about TV shows that dealt only with factual information (plot, cast). At the same time, when it comes to product reviews, texts that contained a mixture of positive and negative sentiment were perceived as neutral.

them (e.g., [53, 156, 131]). The use of words as sentiment markers (features) in sentence- or text-level sentiment annotation systems requires the development of lists of words annotated with sentiment tags. The research on word-level sentiment annotation has produced a number of such lists of words that were manually or automatically tagged with sentiment and related categories. The following sections provide an overview of publicly available manually annotated lists that can be used as input or as a gold standard for automatic annotation, and of a number of automatic methods of word-level sentiment annotation.

2.2.1 Manually Annotated Lists

Interest in annotation of words with sentiment and related categories dates back to the 1960s, when social-science content-analysis research drew attention to such semantic properties of words as their evaluative character, power, etc. The first manually annotated list of words tagged with positive and negative sentiment was created as a part of the **General Inquirer** (GI) content-analysis project [118]. The aim of the project was to code and classify texts using content-analysis methods and the information contained in three main word lists: (1) Harvard IV-4 dictionary of content-analysis categories that include Osgood's three semantic dimensions (value, power and activity); (2) Lasswell's dictionary developed by Lasswell and Namewirth [91]; and (3) five categories based on social cognition work of [112].

The "value" category in the latest version of the Harvard IV-4 list is the most relevant for sentiment analysis. It consists of two annotations — "Positiv" (assigned to 1915 words) and "Negativ" (assigned to 2291 words). The remaining 5669 GI entries were not assigned to any of the two categories and can be considered unmarked, or neutral.

The General Inquirer list is widely recognized as a gold standard for evaluation of automatically produced lists of words annotated with sentiment. The presence of the third, neutral, category makes GI a particularly valuable resource for evaluation of word-level sentiment tagging systems because it enables the evaluation of annotation accuracy not only in binary (positive-negative), but also in ternary (positive-neutral-negative) classification. The separation of neutrals from sentiment-bearing (positive or negative) words, as mentioned earlier, is a much harder task both for automatic systems and for human annotators.

A recent large-scale effort in tagging an extensive list of English words with affect tags was done semi-automatically by Strapparava and colleagues [122, 134] as part of the WordNet Domains project [77]. They created an addendum to WordNet [42] named WordNetAffect, where they assigned a number of affect labels to words in WordNet and then expanded the lists of labeled words using WordNet relations of synonymy, antonymy, entailment, and hyponymy. In addition to a range of semantic labels that are based on psychological and social science theories [92, 37, 36], WordNetAffect also includes some attributes assigned to emotions: *valence* (positive or negative) and *arousal* (the strength of the emotion). The latest (2004) version of WordNetAffect covers 1314 synsets[5] and includes 3340 words.

Other, smaller word lists that can be used in sentiment annotation include Whissell's dictionary of Affect in Language (DAL) [124, 135, 136], Affective Norms for English Words (ANEW) [19], the list of sentiment-bearing adjectives compiled by Hatzivassiloglou and McKeown [52], Linguistic Iquiry and Word Count (LIWC) Dictionary[6], and lists of sentiment and subjectivity clues from work by Wiebe and colleagues[7].

Manually annotated lists are an excellent resource for evaluation of word-level automatic tagging systems and for text-level sentiment classifiers. Nevertheless, there is a number of limitations on the use of manually annotated lists for these purposes: limited coverage [50], low inter-annotator agreement [62, 121], and the diversity of tags used:

Limited coverage. A good quality manual annotation requires extensive time and a focused effort of a team of at least two annotators. Since such annotation is hard to perform on a large scale, the existing manually annotated lists have a fairly limited coverage. Grefenstette et al. [50] compared three manually annotated lists: GI, list from [123] and Clairvoyance Gold standard — a proprietary sentiment lexicon developed by Clairvoyance. The size of intersections between any two of these lists was only 22-23%.

Low inter-annotator agreement. A comparison of several manually annotated lists shows also a surprisingly low inter-annotator agreement among independent teams of annotators. Kim and Hovy [62] report for their experiments with two annotators who agreed only on 76% of positive-negative-neutral labels assigned to adjectives and on 62% of such annotations for verbs. Such consistency in the reported agreement rates suggests that low inter-annotator agreement rates reflect the properties of the category of sentiment, rather than variability in quality of annotation. It should be noted that this low inter-annotator agreement is reported for the most coarse-grained differentiation of positive, negative and neutral words. As more fine-grained manual annotation categories are introduced, the problem of low inter-annotator agreement is likely to be further exacerbated as has been shown by Strapparava and Mihalcea [121].

Diversity of the tags used. While positive and negative tags are relatively straightforward, more fine-grained tags assigned by some manually annotated lists are more difficult to interpret. Since different lists are based on different theories (e.g., affect theories such

[5] According to WordNet glossary, synsets consist of words and collocations that are interchangeable in some contexts and therefore have the same meaning, explained by the gloss and examples.

[6] See http://www.kovcomp.co.uk/wordstat/LIWC.html

[7] See `http://www.cs.pitt.edu/mpqa/~opinionfinderrelease`

as Ortony et al. [92], Ekman's [36] basic emotions, appraisal theory [80], etc.) these fine-grained annotations are not directly comparable across different lists. An interesting attempt to address this problem was done by Strapparava and Valitutti [122], Valitutti et al. [134] in their work on WordNetAffect where they assigned labels based on three different affect theories at the same time: Ortony's et al. [92] classification of terms into emotional, non-emotional affective and non-affective, Elliot's [37] 24 affective categories and Ekman's 6 basic emotions. Yet, as mentioned earlier, inter-annotator reliability of such annotations can be problematic.

2.2.2 Automatic Annotation

The limited coverage, low inter-annotator agreement, and the diversity of tags used in manually annotated word lists prompted the development of automatic systems for word-level sentiment and subjectivity annotation. The methods employed for assigning sentiment and related tags to words can be subdivided into two groups based on the main resource used in such systems: (1) corpus-based approaches, that deduce a word's sentiment from its behavior in texts, and (2) dictionary-based methods, that rely on the information in existing lexical resources (dictionaries and thesauri) in order to infer the sentiment of words and/or their senses.

Corpus-based methods

The 1997 study by Hatzivassiloglou and McKeown pioneered the corpus-based automatic **sentiment tagging at the word level** as a field of research in NLP. Their method builds upon the observation that some linguistic constructs, such as conjunctions, impose constraints on the sentiment of their constituents, as in the following example from [52]:

$$\textit{The tax proposal was} \left\{ \begin{array}{l} \text{simple and well-received} \\ \text{simplistic but well-received} \\ \star\text{simplistic and well-received} \end{array} \right\} \textit{by the public.}$$

The conjunctions collected from the Wall Street Journal 1987 corpus were first organized in a graph using a supervised machine-learning algorithm. At the next step, adjectives were clustered into two sets depending on the type of link between them. The cluster that had higher average frequency was deemed to contain positive adjectives, and the cluster with lower average frequency — negative sentiment, based on the correlations observed in [51]. The algorithm described in [52] is limited, however, to adjectives (and probably adverbs as observed by Turney and Littman [131]) and requires large amounts of labeled data to produce accurate results.

Turney [130, 131, 132] proposed a more general method that does not require annotated data for training. This method induces word sentiment from the strength of its association with 14 seed words with known positive or negative semantic orientation (SO). The method has two main variations: SO-PMI and SO-LSA. In **SO-PMI**, the association of a word with positive and negative seed words is established by computing the point-wise mutual information (PMI) [27]. For each word, the system ran 14 queries on AltaVista using the NEAR operator to acquire co-occurrence statistics of this word with the 14 seed words. Another query retrieved frequency statistics for single words. The combinations with positive seed words resulted in positive scores, and the pairs with negative seed words produced negative scores. The scores for all 14 pairs were then combined into a single score that was interpreted as a measure of the association of a word with a positive or negative category. The results of the system runs using the SO-PMI method were evaluated against GI on a variety of test settings and gave up to 97.11% accuracy (when only the top 25% of the words with highest confidence were classified, for the full test set the accuracy was 82.84%). The size of the corpus had a considerable effect on the quality of the resulting lists: the use of a 10-million word corpus instead of the full Word Wide Web content reduced the accuracy of the method to 61.26-68.74%.

The second approach used by Turney and Littman [131] — **SO-LSA**, calculated the strength of semantic association between words using Latent Semantic Analysis (LSA) [72] techniques[8] and the largest available corpus, made of 10-million words. On a corpus of this size, SO-LSA performed relatively better than SO-PMI (on a ten-million corpus it can give up to 81.65% accuracy while SO-PMI on the same corpus gives at most 68.74%), but SO-LSA is much more complex and is harder to implement.

Due to its simplicity, high accuracy and domain independence, the SO-PMI method became very popular (see, for example, [13, 126, 46]) until AltaVista discontinued the support for the NEAR operator in 2005, and no further experiments using this method on the WWW corpus could be conducted. The attempts to substitute NEAR with the AND operator lead to considerable deterioration in system performance [132, 13, 125].

In **word-level subjectivity analysis**, Bethard et al. [14] proposed a way of making use of co-occurrence information for the acquisition of opinion words (e.g., *accuse, disapproval, belief, commitment*) from texts. They used two different methods. In the first method, they calculated the frequency of co-occurrence of words from the corpus in the same sentence with seed words which were taken from [52], and computed the log-likelihood ratio using this information. The second method tested by Bethard et al. [14] computes relative frequencies of words in subjective and objective documents. The first method produced better results

[8] see http://lsa.colorado.edu

for adverbs and nouns (P: 0.90, R: 0.38) and gave a higher precision but lower recall for adjectives (P: 0.58, R: 0.47). The second method worked best for verbs (P: 0.78, R: 0.18).

Similarly, Kim and Hovy [63] separated opinion words from non-opinion words by computing their relative frequency in subjective (Editorial) and objective (nonEditorial) texts from TREC-03 Novelty Track data [115]. The relative frequency scores reflected the bias of each of the words toward editorial or non-editorial texts. The final list consisted of 15,568 words. The list was not compared to any gold standard.

Riloff and Wiebe [105] and Grefenstette et al. [50] use syntactic patterns for detection of word subjectivity. [105] applied two bootstrapping algorithms — Basilisk [128] and MetaBoot [103] — to learn from a corpus lexico-syntactic expressions that are characteristic for subjective nouns. The algorithms are fairly accurate in the beginning of the process (at 20 words, the accuracy is 95%) but their performance deteriorates significantly after multiple bootstrapping iterations (after 1000 words, MetaBot has only 28% accuracy, while the performance of Basilisk is 53%).

Wiebe et al. [140] report experiments on automatic tagging of single-word direct opinion expressions (*ons*)[9] using two classifiers trained on annotated texts: k-nearest-neighbors (k-NN) and Naïve Bayes. Features used in the experiments included part-of-speech of the target word and information about its immediate context: two nearest words, their part-of-speech and dependency parse chunk. In these experiments, K-NN had P: 0.70, R: 0.63; Naïve Bayes — P: 0.47 and R: 0.77. The precision of these methods is low compared to the performance of automatic approaches to sentiment tagging that seek to distinguish positive and negative words but is comparable to other methods that differentiate 'opinion' (or subjective) words from objective (i.e., neutral) ones like [14].

Corpus-based approaches can produce lists of positive and negative words with relatively high accuracy (up to 97%). The majority of these methods, however, require very large annotated training datasets to realize their full potential. Some of the limitations of corpus-based approaches are overcome by the dictionary-based methods that rely on existing lexicographical resources (such as WordNet) as a source of semantic information about individual words and senses.

Dictionary-based methods

Dictionary-based methods make use of the information contained in existing lexicographical and reference resources, such as WordNet and thesauri, in order to assign sentiment to a

[9] The term *ons* is used by [145] as a general term for direct expressions of potential opinions such as *speech events* and *private states*, e.g. disapproval, interest, etc.

large number of words. Most such methods use the thesaural relations between words (synonymy, antonymy, hyponymy/hyperonymy) to find the similarity between the seed words and other entries (e.g., [60, 62]). Several recent approaches also exploit the information contained in dictionary definitions in word-level sentiment tagging [58, 38, 39].

As opposed to corpus-based methods, dictionary-based methods are often applied not only to two-way (positive vs. negative) classification, but also to three-way, positive vs. negative vs. neutral, categorization. Dictionary-based methods can assign sentiment not only to words, but also to their senses, since the labels are based on sense-level definitions. Thus, one of the main advantages of these methods is their applicability to sentiment tagging not only at the word-level but also at the sense level, which is a new and promising direction in sentiment tagging research [142].

The first attempt to employ WordNet relations in word sentiment annotation was made by Kim and Hovy [62, 63], who proposed to extend lists of manually tagged positive and negative words by adding their synonyms to the list. Starting from 54 verbs and 34 adjectives, they applied the method in two iterations and acquired 12113 adjectives and 6079 verbs. The acquired words were then ranked based on the strength of sentiment polarity assigned to each word. This strength-of-sentiment score was computed by maximizing the probability of the word's sentiment category given its synonyms. The accuracy of the approach in case of the "lenient agreement", when neutrals were collapsed with positives, was 68% for adjectives and 73-76% for verbs. With bigger seed lists, the agreement went up to 76-78% for adjectives and to 79-81% for verbs. A similar approach was used by Hu and Liu [56], who used synonymy relations to extract opinion-words from WordNet.

An alternative way of making use of WordNet synonymy relations for tagging words with Osgood's three semantic dimensions was proposed by Kamps et al. [60]. In order to assign positive or negative values to a word, they computed the shortest path connecting that word to the words *good* and *bad* through WordNet relations, such as synonymy and antonymy. The authors report the accuracy of 68% compared to the original set of evaluative adjectives in GI and 67.32% accuracy compared to the extended version of the evaluative category. The neutral adjectives from GI were included in the intersection on which the evaluation was performed.

Dictionary-based approaches to word-level sentiment tagging do not require large corpora or search engines with special functionalities. Instead, they rely on generally available existing lexical resources like WordNet. They can produce accurate, comprehensive, and domain-independent lists of words and/or their senses annotated for sentiment and subjectivity. Such lists constitute an important resource for sentence or text sentiment tagging and, since they are compiled in advance, they can improve efficiency of sentiment tagging at sentence and text level. In Section 4.1, I describe a new, accurate approach to fine-grained word level sentiment tagging using WordNet-derived word lists.

2.3 Sentence and Text-level Sentiment and Subjectivity Analysis

The ultimate goal of sentiment and subjectivity annotation is the analysis of clauses, sentences, and texts. The acquisition of individual sentiment-laden words/unigrams using corpus-based or dictionary-based methods usually serves as a critical first step in the analysis of such larger linguistic units. In this section, I review the role of words/unigrams and other features and the approaches used in automatic sentiment/subjectivity tagging at the text and sentence level and also assess the corpora available for training and testing of such systems.

2.3.1 Manually Annotated Corpora

Figure 4 (p. 35) describes manually annotated corpora that are publicly available at the present time (2008)[10].

The scarcity of manually annotated resources for system training and evaluation poses substantial challenges for sentiment/subjectivity research. As a result, researchers often resort to user-created rankings in online product, book, or movie reviews as a proxy for professional quality sentiment/subjectivity annotation (Figure 5 p. 36). Such textual data that usually comes with some ranking scale (good-bad, liked-disliked, etc.) is easily available for some genres and is fast to collect. Nevertheless, without manual screening and cleaning, the corpora obtained this way (e.g., popular Cornell movie-reviews datasets [94, 95]) contain a significant amount of noise (e.g., erroneous ratings that do not correspond to the sentiment expressed by the review, misspellings, phrases in different languages, etc.). The problem of corpus quality is exacerbated by researchers breaking these texts automatically into sentences or snippets in order to create datasets for sentence- or snippet-level annotation, since the sentiment assigned to snippets is extrapolated from the overall sentiment of the review. Thus it takes for granted the correctness of the review sentiment label and expects that sentences that make the review will have the same sentiment as the text overall, which is not always true.

[10] Below are listed only publicly available free datasets that were relevant for this dissertation. A more complete annotated list of available corpora can be found in [93].

2.3.2 Inter-Annotator Agreement Studies

The importance of performance results reported for an automatic annotation system cannot be meaningfully assessed without comparison to human inter-annotator agreement on a given corpus or genre of texts (newspaper texts, parliamentary hearings, reviews, blogs, etc.). The rate of agreement among human annotators may reveal important insights about the task and can provide a critical baseline for system evaluation. Unfortunately, only a few inter-annotator agreement studies have been conducted to date [146, 64, 62, 121] and inter-annotator agreement on some types of corpora popular in sentiment research, such as movie reviews and blogs, remains completely unexplored. In the few inter-annotator agreement studies conducted to date, it can be observed that the agreement depends on the unit that is annotated (sentence vs. text), on the annotation type (subjectivity vs. sentiment), on the domain, the genre, and on some other factors.

The available inter-annotator agreement studies for **sentence-level annotation** display a substantial variability in results. In a study on **sentence subjectivity** labels, Bruce, Wiebe and O'Hara [149] instructed annotators to classify a sentence as subjective if it contained any significant expression of subjectivity. After multiple rounds of training and annotation instructions adjustment, the pairwise Kappa[11] over the entire test set from Wall Street Journal (WSJ) ranged from 0.59 to 0.76. Another study reported in [150, 146] was conducted on the WSJ data with a more detailed annotation schema at the private state[12] expression (word or phrase) level produced considerable higher agreement: the average κ=0.77 (κ ranged from 0.72 to 0.84 and overall agreement ranged from 88% to 94%). The authors concluded that the improvement in the agreement "suggests that adding detail to the annotation task can help annotators perform more reliably" [146]. The presence of additional labels (topic and holder) may have contributed to the difference in the results reported in the two studies of sentence-level subjectivity annotation conducted by Kim and Hovy. In the first study [64], where annotators assigned only opinion/non-opinion labels to sentences, the agreement on 100 sentences was 73% (κ=0.49), while in the second study [66], where 100 sentences were annotated also with opinion holder and topic, the agreement was 82% (κ not reported).

The agreement on **sentence-level** sentiment annotations was explored in two studies on two different domains and showed substantial variability in results. Kim and Hovy

[11] Cohen's kappa [28] measures the agreement between two annotators who classify items into mutually exclusive categories, relative to the "chance agreement". Higher values of kappa indicate stronger agreement, and $\kappa < 0.85$ is interpreted as near perfect agreement.

[12] Wilson and Wiebe [150] define *private states* as a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments. A private state is a state that is not open to objective observation or verification.

[62] report relatively high agreement (κ=0.91) between two annotators who independently assigned positive, negative, and n/a labels to 100 newspaper sentences from DUC 2001. At the same time, the inter-annotator agreement reported by Gamon and Aue [46] on a corpus of car reviews produced the pairwise Kappa of only 0.70 — 0.80.

A study by Mihalcea and Strapparava [121] suggests that inter-annotator agreement is substantially lower on more fine-grained types of annotation, i.e., when judges are required to distinguish among several related categories. In their study conducted for the new semEval Affective Text task, six annotators assigned sentiment score and Eckman's [36] six basic emotions scores to news headlines. The agreement, as measured by Pearson correlation, was 78.01 for the category of sentiment, while Pearson correlation for the emotion labels ranged between 36.07 (for surprise) to 68.19 (for sadness). Such a low level of inter-annotator agreement suggests that fine-grained emotion annotation is very subjective and the categories have to be defined more precisely before the task can be meaningfully treated by the automatic methods. The task is further complicated by the low number of examples for some emotions (e.g., *disgust* was present only in 41% of the headlines) and by interactions between emotions (i.e., there were only 12 examples of headlines with a single emotion in the corpus).

Overall, the inter-annotator agreement studies on news texts are scarce and inconclusive: the only study of agreement in sentiment tagging of news, reported by Kim and Hovy [62] suggests that the task is relatively easy to humans, at the same time, the agreement in subjectivity analysis of news sentences was lower, and kappa varied from 0.49 to 0.84 (the agreement was from 73% to 94%). These numbers do not allow the inference of any reliable conclusions about the complexity of the task and have to be confirmed by more research on other datasets.

Human inter-annotator agreement studies provide an important baseline that sets performance expectations for automatic tagging systems. The lack of such studies for some popular domains (e.g., movie and product reviews) imposes serious limitations on the rigor of system performance evaluations. Further studies are also needed to explore the factors that affect the rates of inter-annotator agreement and, hence, are likely to affect the performance metrics of automatic sentiment annotation systems. These factors may include genre of the text, granularity of the category (binary sentiment vs. six emotions), and unit size (e.g., text vs. sentence).

2.3.3 Automatic Tagging at the Sentence and Phrase Level

Since a single text often includes both positive and negative sentences, more fine-grained linguistic units — sentences and clauses — are often regarded as the most natural classification units for sentiment/subjectivity annotation [141]. The sentiment of an individual sentence is usually more homogenous than that of the whole text, but is harder to identify due to a very limited number of subjectivity/sentiment clues on which the sentiment attribution decision has to be made by the system. For this reason, sentence-level tagging would tend to produce higher precision but lower recall than text-level annotation [57]. These properties of text and sentence level annotation have stimulated attempts to improve system accuracy by performing annotation simultaneously at different levels [82, 114, 79].

At sentence level, most of the *subjectivity annotation* research is conducted on the MPQA corpus [140], which provides a reliable gold standard for training and evaluation of subjectivity annotations of newspaper texts at this level. In *sentiment tagging*, however, researchers have to rely on small, manually annotated test sets (e.g., [62, 56]) created *ad hoc* for a given study. Such sets are seldom made public after the experiments are done and are usually too small for a meaningful application of machine learning methods. The two major factors that influence the performance of sentence-level classifiers are the features and the algorithm that are used for classification.

The use of words/unigrams provides a hard-to-surpass accuracy in sentence-level sentiment tagging [95, 56]. Kim and Hovy [62] found that simple presence of sentiment-bearing words worked better than more sophisticated scoring methods. The results of several independent experiments, however, suggest that some improvement can be gained if multiple feature sets are combined. In **subjectivity tagging**, Hatzivassiloglou and Wiebe [53] have shown that a combination of lists of adjectives tagged with dynamic, polarity, and gradability labels was the best predictor of sentence subjectivity. Riloff and Wiebe [105] compared the results of 25-fold cross-validation experiments with Naïve Bayes on various sets of features. The accuracy of the system consistently improved with addition of every new feature: it went from 72.1% in experiments with a predetermined set of subjectivity clues from [149] to 74.3% when subjective nouns obtained using extraction patterns were added, and then to 76% when supplementary manually constructed features and quantification of clues density in the immediate context were added. The experiments conducted by Breck et al. [20] confirm that the use of large set of diverse features (lexical, syntactic, dictionary-based) improves the opinion expression identification accuracy. Similarly, in the experiments on **sentiment tagging** conducted by Aue and Gamon [11], the increase in the number of unigram features obtained by generating a larger list of sentiment-bearing words resulted in an increase in the average precision (from 44.8% to 49.57%) and recall: (from 45.1% to 49.95%)[13].

The expansion of the feature set through a combination of different kinds of features, however, does not necessarily produce a marked improvement in accuracy. In experiments on sentence-level subjectivity detection on the MPQA corpus data, Riloff et al. [104] obtained very close results for unigrams taken alone and for unigrams in combination with bigrams or with extraction patterns. Similarly, sentiment annotation experiments by Yu and Hatzivassiloglou [156] also did not show any significant improvement when bigrams or trigrams where added to the feature set. These results suggest that, for sentence-level annotation, unigrams are the most useful type of feature (as opposed to the text-level annotation where higher order n-grams can improve the accuracy of sentiment tagging). Further research is needed in order to confirm this hypothesis (see experiments described in Chapter 3).

Some other features have also been successfully used in sentence-level annotation. In the task of subjectivity classification, these include a presence of complex adjectival phrases (ADJP) [14], similarity scores [156], and position in the paragraph [149]. In sentiment determination, performance gains were also associated with the use of syntactic patterns and incorporation of negation [57, 89, 7], as well as the use of knowledge about holder [62] and target of the sentiment [56].

Classifier algorithms

The studies on sentence level sentiment and subjectivity annotation suggest that these two tasks require distinct approaches for best accuracy.

In sentence-level **subjectivity tagging**, statistical approaches that use Naïve Bayes or Support Vector Machines (SVM) significantly outperform the non-statistical techniques. The best reported accuracy of non-statistical classifiers used in this task reached only 67.3% for objective clauses and 71% for subjective clauses detection [143]. The experiments by Riloff et al. [104] on a comparable set of newspaper texts produced the best accuracy of 74.3% for Naïve Bayes when adjectives and nouns were used as features [106], while SVM with unigrams reached 74.8% accuracy [104][14].

In the experiments reported in the extant literature for a binary (positive-negative)

[13] Aue and Gamon [11] report average precision and recall across theree class labels, positive, negative and neutrals.

[14] On a fairly different and not directly comparable set of movie review snippets, Pang and Lee [95] report 90% accuracy in subjective/objective classification of sentences with SVM and 92% with Naïve Bayes. It should be noted that movie reviews are known to yield much higher accuracy in sentiment annotation than news texts. This difference is likely due to the size and uniformity of the movie reviews corpus from [95].

sentiment classification of sentences, non-statistical methods that rely on a simple presence of sentiment markers in a sentence or on the strength of the sentiment associated with these markers yield better accuracy than statistical approaches. This is probably due to the lack of annotated training data. The best accuracy (84%) was reported for a classifier that used the average number of opinion words in a product review sentence in order to determine its sentiment [56]. The ties in this study were broken using the word's proximity to the sentiment target as an additional clue. Aue and Gamon [11] conducted experiments on sentence-level sentiment classification of product (car) reviews using several classifiers. Their results permit to assess the relative performance of different classifiers on the same data. The difference between SVM and Naïve Bayes was very small: average precision with SVM (trained on 2000 examples) was 52.97% and average recall 52.56% compared to 51.61% precision and 51.05% recall for Naïve Bayes. A non-statistical classifier performed slightly worse: its precision was 49.57% and recall 49.95%.

Overall, the results reported in recent literature suggest that both statistical and non-statistical (lexicon-based) methods have certain advantages in sentiment and subjectivity classification at the sentence level. However, to date very little research has been conducted into sentence-level subjectivity and sentiment and in the few studies reported in the literature, the results are not directly comparable with each other because of diverse datasets and approaches used. Further research is required to explore the benefits and limitations of these approaches. Chapters 3 and 4 will address some of these questions.

2.3.4 Automatic Tagging at the Text Level

Features

The choice of features used by a sentiment annotation system is a critical factor affecting its performance. A wide range of features has been used in sentiment and subjectivity analysis systems: lists of words, lemmas or unigrams, bigrams, and sometimes higher-order n-grams, parts-of-speech, syntactical properties of surrounding context, and other.

While unigrams are sometimes considered to be a sufficient feature for determination of sentiment or subjectivity of a text [94], experiments by Dave et al. [32] and Cui et al [30] demonstrated that with sufficiently large[15] training sets, higher order n-grams perform better than lower-order ones. For sentiment classification, the combination of n-grams of different order yields the best performance (up to 90.5% accuracy on movie reviews [11]), especially when Expectation Maximization-based feature selection was applied [11]. In

[15] Dave et al. [32] evaluated the performance of uni-, bi-, and tri-gram models on product review texts. They used a training corpus of 2700 to 5800 examples, depending on the product and type of experiment. Cui et al. [30] assessed the performance of one- to six-gram classifiers using 200,000 online reviews texts.

subjectivity tagging, Wiebe et al. [144] also obtained a better accuracy when unigrams (adjectives and verbs) were combined with n-grams. Adding bigrams to unigrams slightly improved the accuracy of both sentiment and subjectivity detection in the experiments reported by [104]. These observations suggest that use of multiple models can improve the performance of both sentiment and subjectivity classifiers. Overall, the use of different features within the same classifier [144, 53, 104, 45] or in a community of several classifiers [31] has a positive effect on system performance for both tasks. The experiments where standard n-gram models or word lists were augmented with context information confirm that these extra features provide additional information to the classifier and contribute to its accuracy. For instance, Gamon [45] observed that a combination of several features (including n-grams, part-of-speech n-grams, frequencies of function words, and a number of features associated with syntactic structure (phrase structure patterns for parse tree constituents, part of speech information coupled with semantic relations, and logical form features such as transitivity of a predicate or tense)), performed better than any of the feature types in isolation. Other additional features, that improve sentiment classifier performance include membership in a collocation [144] and a combination of words annotated for different semantic categories related to sentiment or subjectivity [137, 43, 90].

Given the number of possible features that can be learned by a classifier, using all features in combination can be costly or even computationally infeasible. For this reason, the selection of the most useful features associated with the greatest gains in performance becomes an important research question in sentiment and subjectivity research. Gamon [45] demonstrated the effectiveness of feature selection based on log-likelihood ratio computed with respect to the target variable [34] for sentiment tagging of product feedback messages. Aue and Gamon [11] report best accuracy when n-grams were selected based on the Expectation Maximization algorithm. Riloff et al. [104] addressed the issue of feature selection by exploring the impact of subsumption hierarchies[16] for different types of features and their relations to each other. They applied two kinds of features: n-grams (unigrams and bigrams) and extraction patterns (EP)[17]. The subsumption hierarchy was used to reduce the feature set: if a feature's words and dependencies were a subset of a more general ancestor in the subsumption hierarchy, the feature was discarded. In addition, the quality of each feature was estimated using information gain and only features with higher information gain were

[16] Riloff et al. [104] uses the following definition of subsumption: "We will say that feature A *representationally subsumes* feature B if the set of text spans that match feature A is a superset of the set of text spans that match feature B. For example, the unigram *happy* subsumes the bigram *very happy* because the set of text spans that match *happy* includes the text spans that match *very happy*".

[17] *Extraction patterns* are patterns that represent role relationships in noun and verb phrases. In [105] they are used to represent subjective expressions that have non-compositional meaning like *drive somebody up the wall.*

allowed to subsume less informative ones. The authors discovered that for both sentiment and subjectivity analysis, a combined use of subsumption and traditional feature selection improves performance of subjective/objective and positive/negative text-level classification by 1–2%.

Feature selection based on linguistic characteristics, such as part-of-speech, also has an impact on system performance. In subjectivity tagging, adjectives were shown to be the best predictors of subjectivity [53]. Other parts of speech (modals, pronouns, adverbs and cardinal numbers) have also been successfully used as subjectivity clues [149, 22]. In sentiment tagging, however, a combined use of words from all parts-of-speech produced more accurate tags: Blair, Salvetti et al. [15, 111] reported 75.5% sentiment annotation accuracy with adjectives only vs. 79.5% accuracy with all POS included, [94] achieved 77% accuracy for adjectives and 80.3% for the same number of unigrams drawn from all parts-of-speech[18].

The frequency of unigrams can also indicate how useful they will be for sentiment analysis. Less frequent unigrams provide more information than more frequent ones. For instance, Pang and Lee [94] and Wiebe et al. [144] observed that neologisms and *hapax legomena* — unique words, observed only once — were among the best sentiment predictors.

Feature generation

Most sentiment classifiers use standard machine learning techniques to learn and select features from labeled corpora. Such approaches work well in situations where large labeled corpora are available for training and validation (e.g., movie reviews), but they fall short when training data is scarce or comes from a different domain [11, 102], topic [102], or time period [102]. The experiments conducted by [102] confirmed that the more homogenous the corpus, the better the performance: when movie reviews used for training and testing were collected from the same web-site with a one-year interval, the performance of the SVM classifier dropped compared to the experiment where both datasets came from the same year.

The realization of limitations of supervised machine learning brought about an increased interest in unsupervised and semi-supervised approaches to feature generation. Aue and Gamon [11] have shown that systems trained on a small number of labeled examples and large quantities of unlabelled in-domain data perform relatively well even in comparison to training on large amounts of in-domain examples (e.g., on the knowledge base feedback corpus, the method that relied on the unlabelled data produced 73.86% accuracy compared

[18] Pang and Lee [94] set a cut-off at 2633 most frequent features for both lists.

to 77.34% for in-domain training and 72.39% for best out-of-domain training experiment).

Blitzer et al. [16] applied structural correspondence learning [17] to the task of domain adaptation for sentiment classification of product reviews. They showed that, depending on the domain, a small number (e.g., 50) of labeled examples can be sufficient for adaptation of the model to a new domain. The authors noted, however, that the success of such adaptation and the number of necessary in-domain examples require substantial similarity between the two domains (e.g., a model trained on electronics performed well on reviews of kitchen appliances, but not on book reviews).

Goldberg and Zhu [49] applied a semi-supervised learning algorithm to movie review classification: they created graphs both on labeled and unlabelled data and then used linear ϵ-insensitive support vector regression to generate a rating function over the combined graph. The accuracy of this method, however, is significantly lower than the performance of a supervised classifier trained on a large number of training examples. When the training set was reduced the accuracy went down by 3% (when half of the original data was used for training) to 10% (when training was limited to 100 examples).

Overall, the development of semi-supervised approaches to sentiment tagging is a promising direction for research in this area but so far the performance of such methods is inferior to the performance of supervised approaches and lexicon-based methods that use general word lists. The availability of a variety of manually and automatically generated word lists and sets of sentiment clues makes lexicon-based approaches an attractive alternative to supervised machine learning when labeled data is scarce.

Domain and genre effects on automatic tagging

A variety of domains and genres have been used in the experiments on sentiment tagging: movie, music, book and other entertainment reviews, product reviews, blogs, dream corpus, etc. The choice of the domain on which a sentiment-tagging system is trained and/or tested can have a major impact on its results. Certain properties of texts and sentences in a given domain contribute to the accuracy of the automatic annotation.

For example, in movie reviews, which are particularly popular domain in sentiment research, positive and negative expressions do not necessarily convey the opinion holder's attitude to the movie: for example, the word "evil", which is highly negative in general texts, is used in movie reviews only with respect to characters or plot, and, thus, does not convey any sentiment with respect to the movie itself [130]. For this reason, simple counting of positive and negative clues in movie review texts is not sufficient for accurate determination of their sentiment, and the clues acquired from out-of-domain sources (such as the negativity of "evil" in general English) often fail here. This property distinguishes movie reviews from product reviews, where the sentiment towards the whole product often is the sum of the sentiment towards its parts, components, and attributes [130]. For this reason, the approaches that rely on general, out-of-domain, information (e.g., words, collocations, etc.) perform substantially worse on movie reviews than methods that rely on in-corpus training (66.7% accuracy [130], vs. accuracy of over 85% with in-domain training [94, 11]). On the other hand, the methods that use general word lists often perform better on product reviews: Turney [130] reports 84% accuracy on automotive reviews and 80% on bank reviews compared to only 65.83% on movie reviews. Kennedy and Inkpen [61] obtained 69.3% accuracy on product reviews vs. 66.7% on movies.

Another important factor that can affect system performance on a given domain is the relative frequency of positive and negative sentences or texts in this domain. It has been observed that texts with positive sentiment are easier to classify than negative ones [61, 57, 32, 70, 24]. The possible explanations to this phenomenon are that positive documents are more uniform than negative ones [32] or that positive clues have higher discriminant value [70]. Another explanation is that negative texts are characterized by extensive use of negations and other types of valence shifters [95] that reverse the sentiment conveyed by individual words (e.g., *not good*). To date there were very few experiments with the use of valence shifters in sentiment and subjectivity annotation and their results are inconclusive. Kennedy and Inkpen [61] observed improvement in system accuracy when valence shifters were taken into account, while Dave et al. [32] report the negative impact of the inclusion of the negation into the feature set. The two experiments differ in two significant parameters that do not allow for a direct comparison: first, Kennedy and Inkpen [61] used several kinds of valence shifters (negations, intensifiers, etc.) while Dave et al. [32] limited themselves to *not* and *no*; second, Kennedy and Inkpen [61] used a bag-of-words approach where valence shifters were added to the feature set, while Dave et al. [32] had special features for the combination of a word with negation (e.g., *NOT_good*).

The observed greater difficulty of negative text annotation requires special care in the design of system evaluation sets. The use of balanced evaluation sets with equal number of positive and negative documents has become a standard in sentiment research.

Classification algorithms

A wide variety of classification approaches has been used for subjectivity tagging at the text level. They include simple keyword counting methods with or without scoring [130, 62], rule-based methods [15], SVM [94, 32, 46], Naïve Bayes [94, 32] and other statistical classifiers used alone, sequentially, or as a community of classifiers. The comparison of the results reported for different methods does not yield any definite answer as to which of these methods is the best for the task of sentiment or subjectivity tagging. For example, in experiments conducted by Pang and Lee [94] SVM outperformed Naïve Bayes in classifying movie reviews when unigram presence, combination of unigrams with bigrams or with part-of-speech information were used as features. At the same time, Naïve Bayes performed better when unigram frequencies or bigrams where considered. Dave et al. [32] compared the performance of the two classifiers and obtained considerably better performance with a Naïve Bayes approach than with SVM when both classifiers used unigrams as features in classification of product reviews. At the same time, SVM was a better choice when bigrams were used as features.

2.4 Conclusions

Sentiment and subjectivity analysis has evolved into a strong research stream in NLP research. State-of-the-art systems can reach up to 90% accuracy on certain domains. The main challenge that sentiment and subjectivity research is facing now, is the creation of robust approaches that can ensure system portability across domains and would not depend on the availability of large amounts of in-domain training data. There are two major approaches to this problem: (1) unsupervised *lexicon-based* methods that seek to automatically create reliable and extensive resources such as lists of annotated words and expressions, and (2) the development of semi-supervised *machine-learning* approaches that will maximize the usefulness of the available resources and ensure the domain adaptation with limited sets of in-domain data.

The research described in this dissertation contributes to the solution of this problem by systematically comparing the two major approaches to sentiment tagging at text-level — unsupervised lexicon-based method that uses general word lists as features and supervised corpus-based methods that learn features from training corpora — and demonstrating their relative advantages and limitations on different genres and domains.

The performance of lexicon-based systems is predicated on the availability of accurate and comprehensive lists of words tagged with sentiment and related categories. The development of dictionary-based methods of lexical acquisition is currently one of the most promising directions of research in this area. A highly accurate dictionary-based approach to fine-grained sentiment tagging of words is described in Section 4.1. The list produced by this approach was one of the inputs for domain-independent unsupervised lexicon-based system used for phrase-, sentence- and text-level sentiment annotation. It was further combined with syntactic information and valence shifters in order to improve the accuracy of sentiment classification (Chapter 4). The dissertation also fills the gaps in the analysis of factors that affect the performance of machine-learning methods. The factors such as unit of analysis (sentence or text), classifier (Naïve Bays or SVM), features (uni-, bi- or tri-grams), and domain/genre are explored and compared here. In order to address the problem of domain-adaptation, which is one of the most important issues in sentiment tagging, a new approach that combines the benefits of machine-learning and lexicon-based methods is proposed here. The method is described in detail and its performance is evaluated on news articles, which is one of the most difficult, yet one of most important domains of practical application of automatic sentiment analysis. The results on news will be compared to the algorithm performance on movie reviews, news articles, product reviews (electronics), and personal blogs.

News texts are among the least explored in sentiment analysis. While some research has been done on the MPQA corpus of newspaper texts in subjectivity classification of phrases, sentences, and texts, sentiment analysis of newspaper data so far was limited to the research performed by Kim and Hovy [62, 63] on a small set of news sentences. Ahmad [1] relies on collocations extracted from large corpora of news texts in order to predict the sentiment of financial markets[19]. Devitt and Ahmad [33] used SentiWordNet [39] with a variety of metrics to assign positive and negative sentiment to financial news. They report very low inter-annotator agreement and system performance on this data: the best F-measure was only 0.57.

A closely related genre of news headlines has been a topic of the Senseval-4 Affective Text task [121]. This new SensEval task provided an interesting opportunity for comparison among different systems run on the same dataset of 1000 manually annotated news headlines. The Affective Text task organizers suggest that news headlines are written as attention grabbers and therefore aim at provoking emotion in the reader. This property makes them particularly suitable for testing of sentiment annotation systems. The headlines were annotated by several judges with their sentiment and six Ekman's basic emotions on the scale of $[-100, 100]$. Inter-annotator agreement and system performance on this data were relatively low, compared to the performance of the same approaches on texts.

A comparison of the available sentence-level sentiment annotation results obtained on news sentences to the results obtained on other domains (e.g., product reviews) has shown that news are significantly harder to tag automatically. This may be due to two major reasons: first, very little labeled data is available for training and fine-tuning of supervised

[19] The paper does not provide technical details about the method or evaluation results.

machine learning algorithms, which have been successful in other domains; second, news sentences cover a variety of different domains, have complex structure, and express diverse and complex opinions, which makes them hard to annotate using machine-learning or lexicon-based methods. New approaches have to be developed in order to obtain accuracy on news that will be comparable to the results on other domains.

In the following chapters, both machine-learning and lexicon-based methods will be applied to the sentences from four domains and genres in order to establish the baselines for each corpus and to identify the factors that influence the performance of the classifiers on this data. The dissertation then proceeds with an introduction of a new approach that combines the strengths corpus-based and lexicon-based approaches, and provides description and evaluation of the proposed approach. This dissertation seeks to validate the hypothesis that semi-supervised learning on training data sets that contain more general and more comprehensive linguistic data, such as dictionary entries, are likely to result in more portable sentiment annotation systems compared to the systems that are trained on a specific domain. Thus, it can be expected that statistical (corpus-based) classifiers, if trained on large sets of in-domain data, would outperform lexicon-based systems trained on General English dictionaries, but would be inferior to lexicon-based systems if trained on small sets of in-domain or on out-of-domain data.

Manually Annotated Corpora
Labels assigned by trained annotators

- Corpus: MPQA (Multiple Perspective Question Answering)

  Level of annotation: Phrases and sentences

  Annotation type(s): Private states (such as emotions or opinions and related attributes)

  Corpus size: 535 documents (10 657 sentences)

  Link: www.cs.pitt.edu/∼iebe/mpqa

  References: [139, 140, 150, 119, 151]

- Corpus: Opinion corpus

  Level of annotation: Expressions and sentences

  Annotation type(s): Subjectivity and objectivity

  Corpus size: 2 sets of documents of 500 sentences WSJ Treeback each

  Link: http://www.cs.pitt.edu/∼wiebe/pub4.html

  References: [149, 22]

- Corpus: Product Reviews

  Level of annotation: Product features

  Annotation type(s): Sentiment features

  Corpus size: 500 reviews

  Link: www.cs.uic.edu/∼liub/FSB/FSB.html

  References: [56]

- Corpus: SemEval-07 Task 14 "Affective Text" dataset

  Level of annotation: Headline

  Annotation type(s): Sentiment (on [-100,100] scale) and 6 Eckerman's basic emotions

  Corpus size: 1225 headlines

  Link: www.cse.unt.edu/∼rada/affectivetext/

  References: [121]

Figure 4: Manually annotated corpora.

Automatically Annotated Corpora
Labels derived from ratings assigned by text authors

- Corpus: Cornell movie-reviews datasets (document level)

  Level of annotation: text

  Annotation type(s): document-level sentiment

  Corpus size: 2000 movie reviews

  Link: www.cs.cornell.edu/people/pabo/movie-review-data/

  References: [94]

- Corpus: Cornell movie-reviews datasets (sentence level)

  Level of annotation: sentence (snippet)

  Annotation type(s): sentence-level sentiment or subjectivity

  Corpus size: 10000 movie review snippets

  Link: www.cs.cornell.edu/people/pabo/movie-review-data/

  References: [95]

- Corpus: Blogs06

  Level of annotation: text (blog post)

  Annotation type(s): opinion polarity and relevance

  Corpus size: 25GB

  Link: ir.dcs.ac.uk/test_collections/access_to_data.html

  References: [81]

- Corpus: Multi-domain Sentiment Dataset

  Level of annotation: text

  Annotation type(s): positive and negative sentiment

  Corpus size: 8000 texts (2000 reviews for each of four domains: DVDs, books, kitchen appliances and electronics)

  Link: http://www.cis.upenn.edu/~mdredze/datasets/sentiment/

  References: [16]

Figure 5: Automatically annotated corpora.

Chapter 3

Setting a Baseline: Supervised Corpus-based Approach

As seen in Chapter 2, supervised machine learning has long been a method of choice for sentiment tagging at the text level. Most approaches to sentiment or subjectivity tagging reported in the extant literature use SVM or Naïve Bayes to classify texts into positive, negative and neutral. These approaches have been very successful in situations where large amounts of uniform training data are available. For example, on movie reviews, the best reported accuracy is close to 90% [46]. Such machine learning methods, however, were not thoroughly tested on smaller units, such as sentences, where lexicon-based approaches are more popular. Given their success in some sentiment annotation tasks, supervised machine learning methods can provide an important baseline to which alternative methods can be compared. This chapter starts with the comparison of the performance of the two most widely used classifiers, Naïve Bayes and SVM, on several genres. Then the portability of such methods will be explored in cross-genre sentiment tagging experiments.

The analysis of the literature (Chapter 2) has shown that several factors can influence the performance of machine learning approaches. *Classification algorithm*, *feature generation and selection*, *corpus size*, and *training and testing domains*, have been identified as the most relevant factors determining system performance in sentiment classification. While these factors can have considerable effect on classifier performance, very little research has been done to systematically and rigorously assess their influence.

This chapter continues Chapter 2 in setting the background for the development and evaluations of the novel approaches presented in Chapters 4 and 5. Its goals are two-fold: First, the experiments presented here fill the gap in the literature on sentence-level sentiment classification by exploring the role of the major factors that can influence the performance of corpus-based (supervised machine learning) approach. Second, it sets the baseline for evaluation of other approaches to sentence-level sentiment tagging and compares sentence- and text-level sentiment tagging. This analysis contributes to our understanding of challenges unique to each of the levels of analysis.

3.1 Data

In my experiments I sought to contrast the performance of several approaches to sentiment tagging on news sentences in comparison with other language domains and levels. Therefore several different corpora were used in this study:

- A set of movie review snippets (further: *movie*) from [96]. According to the procedure described in [96], this dataset of 10,662 snippets was collected automatically from `www.rottentomatoes.com` website. Sentences in reviews marked "rotten" were considered negative and snippets from "fresh" reviews were deemed positive. The resulting sets of snippets were cleaned manually. Snippets usually correspond to sentences, but may occasionally span two short, related sentences or, on the contrary, cover only one clause in a long sentence. In order to make the results obtained on this dataset comparable to other domains, a randomly selected subset of 1066 snippets was used in the experiments.

- A balanced corpus of 1200 manually annotated sentences extracted from 83 newspaper texts (further, *news*). The full set of sentences was annotated by one judge. 200 sentences from this corpus (100 positive and 100 negative) were also randomly selected from the corpus for an inter-annotator agreement study and were manually annotated by two independent annotators. The pairwise agreement between annotators was calculated as the percent of same tags divided by the number of sentences with this tag in the gold standard. The pairwise agreement between the three annotators ranged from 92.5 to 95.9% (κ=0.74 and 0.75 respectively) on positive vs. negative tags.

- A set of sentences taken from personal weblogs (further, *blogs*) posted on LiveJournal (http://www.livejournal.com) and on CyberJournalist (http://www.cyberjournalist.com)[1]. This corpus is composed of 1200 sentences (400 sentences with positive, 400 sentences with negative sentiment, and 400 neutral sentences). In order to establish the inter-annotator agreement, two independent annotators were asked to annotate

[1] It was not possible to use existing blogs corpora such as Blog-06 corpus because they use user's mood labels instead of standard sentiment tags and they don't provide sentence level annotations.

200 sentences from this corpus. The agreement between the two annotators on positive vs. negative tags reached 99% (κ=0.97).

- A set of 1200 sentences from product reviews (thereafter *PRs*) extracted from the annotated corpus made available by Bing Liu [56][2].

The data sets used in the experiments are summarized in Table 1.

Movie review texts corpus created by Pang and Lee [95] and a corpus of 1000 news headlines provided by SemEval Affective Text task organizers [121][3]) were used in comparative experiments as well.

	Movies	News	Blogs	PR	Headlines
Text level	2002 texts	267 texts	n/a	n/a	n/a
Sentence level	1066 snippets	1200 sent.	1200 sent.	1200 sent.	1000 headlines

Table 1: Datasets

The main experiments reported in this chapter were conducted on all four sentence-level corpora listed in Table 1. However, some experiments had to be limited to fewer datasets. Thus, the experiments with ternary classification were conducted on blogs and on news datasets, which include not only positive and negative, but also neutral sentences. Both product and movie review datasets did not contain neutral sentences and therefore could not be used to test the ternary classification. Table 1 includes also the information about two additional datasets: movie review texts and news headlines. These two corpora were used in the experiments where I compared the performance of supervised machine learning approaches on language units of different length. These experiments were limited to these two genres because there are no comparable text-level corpora for blogs and product reviews. Finally, movie reviews snippets dataset, due to its large size, provided an opportunity to explore the influence of corpus size on classifier performance. Other corpora were too small to conduct such experiments.

3.2 Factors Affecting Corpus-based Systems' Performance

As mentioned earlier, *classification algorithm*, *feature generation and selection*, *corpus size*, and *training and testing domains* have been identified in the extant literature as the most relevant factors determining system performance in sentiment classification [95, 11, 16, 32, 104]. Since little research has been done to systematically assess their influence, this section

[2] http://www.cs.uic.edu/~liub/FBS/FBS.html

[3] The task is described in more details in Sections 2.3.2 and 2.4.

explores the role of these three major factors in sentence-level sentiment classification. All experiments were conducted with 10-fold cross-validation. For all tests, except those with feature selection, all features learned from training data have been retained and no stop-word list has been used to prune the feature set. This decision is based on the observation that some of the words that are often part of stop lists may be useful markers of sentiment. For example, Pang and Lee [94] observed that punctuation (e.g., exclamation mark) can contribute to classification accuracy. Wiebe et al. [149] and Sokolva [116]) noted that personal pronouns such as *you*, were markers of subjective, emphatic text. Das and Chen [31], Pang and Lee [94], and Wiebe et al. [144] report that rare words (*hapax legomena* and neologisms) are often the best indicators of the presence of sentiment.

3.2.1 System Performance on Texts vs. Sentences

Since the comparison of sentiment annotation system performance on sentences and texts has not been attempted to date, I sought to close this gap in the literature by conducting a set of comparative experiments on data sets of units of different language levels from the same domain/genre. There is only one pair of datasets that would allow such comparison between text and sentence level — two movie review corpora collected by Pang and Lee from the `www.rottentomatoes.com` website[4]. The first corpus consists of 2002 movie review texts (further, movie texts) from [95], the second is a set of 10662 movie review snippets (5331 with positive and 5331 with negative sentiment). The corpus of movie review snippets is described in greater detail in Section 3.1. In the experiments reported in Table 2, the full set of 10662 snippets was used in order to obtain larger amount of training data for bi- and tri- gram models. The results with 10-fold cross-validation are reported in Table 2[5].

	Trained on Texts		Trained on Sent.	
	Tested on Texts	Tested on Sent.	Tested on Texts	Tested on Sent.
1gram	**81.1**	69.0	66.8	**77.4**
2gram	**83.7**	68.6	71.2	**73.9**
3gram	**82.5**	64.1	70.0	**65.4**

Table 2: Accuracy of Naïve Bayes on movie review sentences and texts.

Table 3 compares the accuracy of Naïve Bayes on news sentences and headlines[6]. In

[4] Both corpora can be downloaded from http://www.cs.cornell.edu/People/pabo/movie-review-data/

[5] All results are statistically significant at $\alpha = 0.01$ with two exceptions: the difference between trigrams and bigrams for the system trained and tested on texts is statistically significant at alpha=0.1 and for the system trained on sentences and tested on texts is not statistically significant at $\alpha = 0.01$.

[6] Table 3 includes only the performance of Naïve Bayes with unigram model because the use of higher order n-grams was not justified given the small size of headlines (only 6.5 words on average).

news genre, headlines constitute an important indicator of the context of the entire text, both in terms of the text's topic and in terms of the sentiment of the news coverage. The dataset used here was provided by the organizers of the SensEval Affective Text task [121]. They claim that news headlines are written as attention grabbers and therefore aim at provoking emotion in the reader. This property makes them particularly suitable for testing of sentiment annotation systems. At the same time, headlines are very short, and therefore, the decision about their sentiment often has to be made based on a single sentiment clue. If this clue is not a part of the feature set (lexicon or n-gram model) used by the system, the headline will be misclassified.

	News Sentences	News Headlines
training set	720 news sentences	800 news sentences + 250 headlines
test set	80 news sentences	1000 headlines
unit size	25.6 words	6.5 words
accuracy	59.5	31.2

Table 3: Accuracy of Naïve Bayes on news sentences and headline (unigrams).

The results in Tables 2 and 3 indicate that granularity of analysis (headlines, sentences or texts) has considerable impact on the performance of the sentiment classifier. The experiments on texts and sentences from movie reviews show that classification accuracy on the same system with the same training set is 12 to 15% lower when it is tested on sentences than when it is tested on full-text movie reviews. Thus a lower number of sentiment clues in a sentence makes the decision about its sentiment is less accurate. The tests with news sentences and headlines further support this observation. The experiments with the headlines were conducted as part of the Senseval-4/Semeval-1 Task 14 Affective Text task. In this task, the original training set provided by the organizers, was very small — only 250 headlines. This data was not sufficient for training of a statistical classifier and we augmented it with the 800 news sentences. Thus, the training sets of the two experiments reported in Table 3 share a large part of the data. Testing, however, was done on the sentences and headlines separately. Despite the addition of the sentence-level training data, the Naïve Bayes classifier performed significantly worse on headlines than on sentences. Thus, the difference in the classifier performance cannot be explained by the properties of the training set, which in both experiments was of approximately the same size and composition. This performance gap was primarily due to the difference in the unit of analysis in the test data and the associated difference in the average length of these units — sentences and texts. The similarity of the trend observed on movie reviews and on news supports the observation that smaller units are consistently more difficult to classify. This can be attributed to the differences in unit size and hence, differences in the number of sentiment clues contained in each unit. Movie review texts, for example, are on average 690 words long, while snippets are made of only 21 words. The difference between news sentences and headlines is also significant: average size of a sentence is 25.6 words, while headlines are only 6.5 words long.

3.2.2 Choice of N-gram Size

Table 2 above also provides insights into the n-gram size interaction with the size of the unit of analysis. Consistent with findings in the literature [30, 32, 46], on the large corpus of movie review texts, the in-domain-trained system based solely on unigrams had lower accuracy than the similar system trained on bigrams. But the trigrams fared slightly worse than bigrams. On sentences, however, the pattern is inverse: unigrams performed better than bigrams and trigrams. These results highlight a special property of sentence-level annotation: *greater sensitivity to sparseness* of the model: On texts, classifier error on one particular sentiment marker is often compensated by a number of correctly identified other sentiment clues. Since sentences usually contain a much smaller number of sentiment clues than texts, sentence-level annotation more readily yields errors when a single sentiment clue is incorrectly identified or missed by the system. Due to lower frequency of higher-order n-grams (as opposed to unigrams), higher-order n-gram language models are more sparse, which increases the probability of missing a particular sentiment marker in a sentence (Table 4[7]). Very large training sets are required to overcome this higher n-gram sparseness in sentence-level annotation. This observation suggests that, given the constraints on the size of the available training sets, unigram-based systems may be better suited for sentence-level sentiment annotation.

[7] The results for movie reviews sentences reported in Table 4 are lower than those reported earlier in Table 2 since the dataset is 10 times smaller, which results in less accurate classification. A smaller subset was selected from the full movie reviews snippets dataset in order to make the results comparable to the experiments on other domains, where considerably smaller amounts of data were available.). The statistical significance of the results depends on the genre and size of the n-gram: on product reviews, all results are statistically significant at $\alpha = 0.025$ level; on movie reviews, the difference between Naïve Bayes and SVM is statistically significant at $\alpha = 0.01$ but the significance diminishes as the size of the n-gram increases; on news, only bi-grams produce a statistically significant ($\alpha = 0.01$) difference between the two machine learning methods, while on blogs the difference between SVMs and Naïve Bayes is most pronounced when unigrams are used ($\alpha = 0.025$).

Dataset	Movie	News	Blogs	PRs
Dataset size	1066	800	800	1200
unigrams				
SVM	68.5	61.5	63.85	76.9
NB	60.2	59.5	60.5	74.25
nb features	5410	4544	3615	2832
bigrams				
SVM	59.9	63.2	61.5	75.9
NB	57.0	58.4	59.5	67.8
nb features	16286	14633	15182	12951
trigrams				
SVM	54.3	55.4	52.7	64.4
NB	53.3	57.0	56.0	69.7
nb features	20837	18738	19847	19132

Table 4: Accuracy of unigram, bigram and trigram models across domains.

3.2.3 Feature Generation and Selection

Feature selection can boost classifier performance by selecting the most useful features and discarding the irrelevant or noisy ones. There are two ways to reduce feature space dimensionality: feature selection using machine learning approaches and selection based on general knowledge, such as lexicons. The former method will be applied in this section. Feature selection methods find a subset of features that give the most information and discard the rest. This selection can be based on human assessment or on automatic attribute filtering. In sentiment and subjectivity analysis, feature selection based on frequency [94], Expectation Maximization [11], Information Gain [48], document frequency [48], Chi-square (χ^2) [48, 88] and other techniques have been used. The feature selection experiments reported in the literature indicate that this procedure can improve the classification results for sentiment analysis [94, 11, 48].

For the experiments reported in this section, three methods of feature selection were applied using the Weka package: Chi-square, Information Gain, and Principle Component Analysis.

Chi-square and Information Gain (IG) are commonly used in NLP to select most relevant features [44]. Chi-square measures the independence of two events (class and feature, in the case of classification) by comparing observed frequencies and frequencies that are expected for independent events.

The chi-square statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-square distribution. The number of degrees of freedom is equal to the number of possible outcomes, minus 1. The chi-square is then used to test whether paired observations on two variables are independent of each other.

On the other hand, Information Gain (IG) [101] is used to find features that provide most information about the classes. IG is one of the most popular measures of association in data mining. The information gain can be interpreted as a way to measure the reduction of uncertainty [78] and selects the features that have the highest information gain as the most relevant ones.

Principal component analysis (PCA) is a vector space transform that is often used to reduce the number of dimensions in a data set with minimum loss of information. PCA reveals the internal structure of the data in a way that best explains the variance in the data.

Feature selection allows retaining a smaller number of most relevant features. For instance, Chi-square and Information Gain both kept the following top ten features for news dataset: *but, under, promote, Quebec, wine, workers, quality, concerned, trust, ingredients.*

The impact of feature selection on the performance of corpus-based classifiers was evaluated in experiments on news sentences and on three other datasets: blogs, movie reviews, and product reviews. Tables 5[8] and 6[9] provide a comparison of the impact of different feature selection approaches on classifier results in binary and ternary classification respectively. The accuracy is compared to statistical classification results with the full set of unigrams. The feature selection using all three methods significantly improves classification accuracy.

Corpus	Blogs		News		PRs		Movie reviews	
Classifier	N. Bayes	SVM	N. Bayes	SVM	N. Bayes	SVM	N. Bayes	SVM
Baseline (no sel.)	60.5	63.9	59.5	61.5	74.3	76.9	60.6	68.6
Chi-square	62.6	**72.1**	65.8	**76.6**	76.8	**82.7**	64.3	**75.0**
IG	62.6	**72.1**	65.8	**76.6**	76.8	**82.7**	64.3	**75.0**
PCA	64.4	68.2	66.7	69.4	74.7	77.3	65.5	71.4

Table 5: Feature selection for statistical classifiers. Binary classification.

As Tables 5 and 6 show, feature selection by all three methods — Chi-square, Information Gain, and PCA — improves classifier performance. For binary classification of News, Naïve Bayes with PCA was 7% better than baseline, while for SVM Chi-square and IG added 15% to the classifier accuracy. For ternary classification of News, the results were

[8] The difference in accuracy between runs with feature selection and baseline is statistically significant at $\alpha = 0.01$ for the majority of the results. However, it is not statistically significant for PCA on News. $\alpha = 0.1$ for Naïve Bayes with χ^2 on Blogs and 0.025 for Naïve Bayes with PCA on Blogs and SVM with PCA on Movies.

[9] For ternary classification of Blogs, the difference in accuracy between results with feature selection and the baseline is statistically significant at $\alpha = 0.01$ for SVM, but not statistically significant for Naïve Bayes. For News, all results are statistically significant at $\alpha = 0.01$.

Corpus	Blogs		News	
Classifier	Naïve Bayes	SVM	Naïve Bayes	SVM
Baseline (no selection)	47.6	51.0	47.0	52.1
Chi-square	47.7	**56.9**	60.5	62.9
IG	47.7	**56.9**	50.6	59.2
PCA	**49.7**	55.2	**64.4**	**68.2**

Table 6: Feature selection for statistical classifiers. Ternary classification.

mixed: Chi-square and IG reduced the accuracy of both Naïve Bayes and SVM, while PCA still brought some improvement (4-5%). For other datasets both in ternary and binary classification all three feature-selection methods improved the performance of classifiers. Similarly to the results on News, the improvement in ternary classification of Blogs was small, especially for Naïve Bayes. In binary classification, the best performance for all genres/domains was obtained with SVM and Chi-square.

It has to be noted that features can also be selected based on general knowledge. The general-knowledge approach to feature selection will be explored in Chapter 4.

3.2.4 Classifier Choice

Two main classification algorithms have been used in sentiment and subjectivity classification: support vector machines (SVM) and Naïve Bayes. These algorithms have a similar performance in sentiment classification at the text level (see [95]). To our knowledge, there were no comparative studies of the two algorithms on sentences. In the study presented here, experiments with both classifiers were conducted using the same data mining package — Weka [152] — with default settings in order to evaluate the impact of algorithm choice on sentence-level sentiment classification results in a controlled setting. The default Naïve Bayes uses Normal Distribution to estimate numeric values and Laplace smoothing for sparse data. The SVM in Weka is implemented using the sequential minimal Optimization (SMO) algorithm [98] and a linear kernel. In these experiments, feature selection was applied in order to assess the impact of classifier choice without influence of other factors.

Tables 8[11] and 7 demonstrate that in sentence-level sentiment classification that uses unigrams as features, SVM performs better on all four genres, both in ternary (for Blogs and News) and binary (for all four corpora) classification. However, it has to be noted that SVM is considerably more demanding on computing resources and runs much slower than

[10] NB stands for Naïve Bayes.

[11] The difference in accuracy between uni-, bi-, and tri-grams is not statistically significant for Naïve Bayes on News, but is statistically significant at $\alpha = 0.01$ for ternary classification of Blogs and at $\alpha = 0.025$ for tri-grams vs. uni-grams for SVM on News.

Corpus	Movie		News		Blogs		PRs	
Num of instances	1066		800		800		1200	
Classifier	SVM	NB[10]	SVM	NB	SVM	NB	SVM	NB
unigrams								
Num of attributes	5410	5410	4544	4544	3615	3615	2832	2832
Accuracy	68.5	60.2	61.5	59.5	63.8	60.5	76.9	74.2
bigrams								
Num of attributes	16286	16286	14633	14633	15182	15182	12951	12951
Accuracy	59.9	57.0	63.2	58.4	61.5	59.5	75.9	67.8
trigrams								
Num of attributes	20837	20837	18738	18738	19847	19847	19132	19132
Accuracy	54.3	53.3	55.4	57.0	52.7	56.0	64.4	69.7

Table 7: N-gram size impact: Sentence level, binary classification (baseline: 50%).

Corpus	Blogs		News	
Num of instances	1200		1200	
Classifier	SVM	Naïve Bayes	SVM	Naïve Bayes
unigrams				
Num of attributes	4575	4575	5644	5644
Accuracy	51.0	47.6	52.14	47.0
bigrams				
Num of attributes	15180	15180	20544	20544
Accuracy	46.6	44.4	50.9	45.6
trigrams				
Num of attributes	19845	19845	27370	27370
Accuracy	36.1	39.7	n/a	44.0

Table 8: N-gram size impact: Sentence level, ternary classification (baseline: 33%).

Naïve Bayes, especially with a large feature space or a large number of instances.

3.2.5 Corpus Size

One of the factors that can influence the performance of a statistical classifier is the size of the training corpus. The most successful text-level sentiment classification approaches have all been tested with large — 2000 and more texts — datasets [95, 11]. [49] have shown that classifier performance on movie reviews tends to deteriorate when the amount of training data is reduced (cutting the training set in half resulted in accuracy that was lower by 3%). At the same time, obtaining sufficient amounts of labeled training data can pose considerable problems in real-life applications, especially for such domains/genres as news, where no author-assigned sentiment ratings exist. This subsection explores the role of the dataset size in the sentence-level classifier performance in a series of tests that were

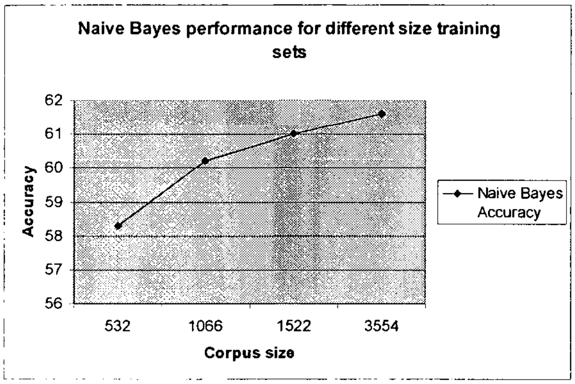

Figure 6: Influence of corpus size on Naïve Bayes performance.

conducted on the same dataset — movie reviews.

In the experiments presented here, a subset of the corpus was randomly selected in order to create corpora of different size. The smallest one was 500 snippets long, while the size of the largest one was determined the maximum size possible with the available computational capacity using Weka package with the same settings: 3554 snippets. The size of the feature set grew proportionately, from 3366 attributes to 11154. As demonstrated by Table 9[12] and graph in Figure 6[13], the increase in corpus size leads to an improvement in accuracy for both Naïve Bayes and SVM. However, this improvement in performance is not dramatic, especially for Naïve Bayes, and the gain in accuracy does not seem to make up for the dramatic decrease in classification efficiency (the bigger the feature space, the more time and memory the process takes). The difference of 2-4% observed on movie reviews sentences when training corpus size went from 1066 to 532 instances is similar to the 3% decrease in accuracy observed by [49] for movie reviews texts. However, further increase in the training set size from 1066 to 1522 sentences did not result in any improvement and even reduced SVM performance.

Training set size	532	1066	1522	3554
Num of attributes	3366	5410	6698	11154
SVM	62.8	68.4	65.6	n/a
Naive Bayes	58.3	60.2	61.0	61. 6

Table 9: Statistical classifiers: Influence of corpus size on binary classification (10-fold cross-validation).

[12] Only the results for SVM are statistically significant at $\alpha = 0.01$.

[13] The graph reflects only the results for SVM, because the difference in the accuracy for Naïve Bayes is not statistically significant.

3.2.6 System Performance on Different Domains and Genres

In Tables 4 (binary classification) and 7 (ternary classification) presented above, one can observe a significant difference in the performance of both SVMs and Naïve Bayes on different domains and genres even with comparable model size and in-domain training. The product reviews are the easiest to classify, followed by movie reviews sentences. Blogs and news are considerably harder to classify, with accuracy on news 7% below the accuracy on movie reviews and more than 15% lower than on product reviews. This difference in performance suggests that results for two systems run on different domains are not directly comparable and that one should not expect that for a given system, sentiment classification of news sentences will be as accurate as classification of product reviews[14].

The difference between domains and genres become even more apparent when a system trained on one domain/genre is ported to another one. For corpus-based learning that relies on the similarities between training and test data, the use of different corpora for testing and training results in deterioration of the classifier performance even when the data comes from the same domain [102]. When domains are different, the negative effect of this dissimilarity becomes more pronounced [46, 16] and depends on the level of similarity between the target and training domains [16].

In this section the issue of domain portability is further explored by comparing in-domain and out-of-domain training. As it has been shown by Aue and Gamon [11], Read [102], and Blitzer et al. [16], training on data that comes from a different domain, or even from a different time period results in much lower accuracy. Table 10[15] confirms that for sentence-level binary sentiment classification, in-domain training consistently yields superior accuracy than out-of-domain training across all four datasets: movie reviews (Movies), news, blogs, and product reviews (PRs). The numbers for in-domain trained runs are highlighted in bold.

The limited domain portability of the machine learning approaches to sentiment classification became the focus of the sentiment research in the last two years, after the work by Aue and Gamon [11] and Read [102] drew attention to this limitation of supervised machine learning methods. There are two main alternatives to the supervised machine learning that

[14] At the text level, [16] who report 7% difference in the baseline accuracy of sentiment classification of kitchen appliances and book reviews. [11] observed 13% difference in accuracy of sentiment classification for movies and knowledge base feedback data.

[15] Statistical significance of the results depends on the corpora and classifier. All results with SVM were statistically significant at $\alpha = 0.01$. With Naïve Bayes, the difference between in- and out-domain trained classifiers was statistically significant at $\alpha = 0.01$ for Product and Movie Reviews. At the same time, no statistically significant difference was observed when classifiers trained on News or on PRs were tested on News. The difference between a classifier trained on Blogs vs. trained on PRs was statistically significant at $\alpha = 0.025$ when tested on Blogs.

Trained on	Tested on							
	Blogs		News		Prod. reviews		Movie reviews	
	N. Bayes	SVM	N. Bayes	SVM	N. Bayes	SVM	N. Bayes	SVM
Blogs	**60.5**	**63.85**	52.21	49.94	58.6	58.8	51.15	53.68
News	55.78	56.25	**59.5**	**61.54**	56.67	57.42	54.80	55.03
Prod. reviews	56.65	56.25	58.08	55.86	**74.25**	**76.92**	55.11	55.83
Movie reviews	49.19	53.16	53.97	55.23	55.5	60.7	**60.57**	**68.5**

Table 10: Out-of-domain vs. in-domain training. Binary classification accuracy (baseline: 50 %).

can potentially address this problem: (1) semi-supervised learning and (2) lexicon-based methods. The semi-supervised approaches try to make the most of small amounts of in-domain labeled data that can be obtained with relatively little effort by bootstrapping from it. Different versions of approaches from this category have been explored in [46, 49, 16, 127]. Their results show that such approaches are much less accurate than methods that rely on in-domain supervised learning and their applicability depends on other factors such as similarity between the domain from which comes the additional training data, and test domain [16, 127].

Another alternative that I will explore in more detail here is to use general lists of sentiment clues and other domain-independent information to assign sentiment to test data (Chapter 4).

3.3 Conclusions

Chapter 3 presented an exploration of the factors that affect the performance of machine learning methods in sentiment analysis task. The analysis of the extant literature (Chapter 2) and experiments reported here identified several factors that impact classifier performance: *classification algorithm*, *feature generation and selection*, *corpus size*, and whether *training is done on the same domain as testing.*

The experiments conducted on news sentences and four other corpora demonstrated that:

- Shorter language units (e.g., sentences and headlines) are harder to classify than larger language units (texts);
- Large training datasets and, consequently, larger feature sets result in higher accuracy. The difference is the most salient when the training corpus size goes up from 532 to 1066 examples. For larger training sets the gain in performance is less apparent;

- Feature selection improves the performance of a classifier. The improvement was the smallest on blogs and on product reviews, and the greatest on news where PCA gave 6% gain in accuracy with Naïve Bayes and 8% with SVM;

- Domain/genre play a crucial role in determining performance of machine learning approaches in sentiment tagging.

- When training is done on a different domain/genre than testing, the performance of a classifier drops significantly. This conclusion is in line with observations reported by [11, 102] for text-level sentiment classification. It highlights the major shortcoming of standard classification approaches to sentiment analysis — the lack of portability across domains and dependence on the availability of large amounts of labeled in-domain (or even "in-corpus") training data.

I argue that portability of sentiment classifiers across domains can be improved by using a more general lexicon-based approach (Chapter 4) instead of corpus-based (and therefore domain-dependent) learning, by using semi-supervised methods [16, 11], or by combing different approaches into an ensemble classifier (Chapter 5).

Chapter 4

Lexicon-Based Approach

Many of the tasks required for effective sentiment tagging of phrases and texts rely on a list of words annotated with some lexical semantic features. Such lists can be produced from corpora (corpus-based approach) or from dictionaries (lexicon-based approach). As opposed to corpus-based learning methods that acquire features from some annotated in-domain or out-of domain corpora, (see chapter 3), the *lexicon-based approach* to sentiment tagging uses general word lists of sentiment-bearing words learned from dictionaries to classify sentences or texts for sentiment. The lexicon-based approach capitalizes on the comprehensiveness and general nature of dictionaries, such as WordNet [42]: dictionaries contain a comprehensive and domain-independent set of sentiment clues that exist in general English. A system that uses such general data, therefore, should be less sensitive to domain changes than corpus-based learners. It can also be more efficient since the list of features is acquired only once and then no further training is required. This saves both time spent on the classifier training and, most importantly, time and effort needed to obtain large amounts of training data from the same domain as the target domain. The dictionary-based approaches, thus, have several theorized advantages over supervised corpus-based machine learning methods: they rely on existing resources, they do not require large corpora and/or specific search capabilities, and they are domain-independent. At the same time, the general character of lexicons used by these methods limits their performance since they do not adapt to different domains or genres[1].

It is important to note that the development of systems for automatic extraction of word sentiment information from dictionaries and other lexicographic resources represents an important task not only for feature acquisition for subsequent annotation of sentences or texts, but also for semantics and lexicography, where the problem of semantic annotation

[1] Next chapter will present an method that takes these properties of lexicon- and corpus-based approaches into account and improves the performance by capitalizing in the complementarity of these two approaches.

at word and even at sense level presents as a task in itself. The automatic annotation of dictionary words with the categories of positive, negative or neutral sentiment, thus, can be regarded as an instance of automatic *word-level semantic tagging* and represents an important direction in NLP research.

One of the common approaches to sentiment annotation of sentences and texts is based on computing an average sentiment of all the words in a text or sentence combined. In this approach, the decision is based on the difference between the number (or the sum of the scores) of positive vs. negative sentiment clues. The NET value of the score (i.e., positives minus negatives) determines the the overall sentiment of the text or sentence. This method relies on lists of words tagged with positive or negative sentiment. Several manually annotated lists have been produced, such as General Inquirer (GI) [118] and list used by Hatzivassiloglou and McKeown [52] (HM). Strapparava and Valitutti [122] created WordNet Affect, an extension to WordNet manually assigning affect labels based on theories of emotion representation to WordNet synsets. As dicussed in Chapter 2, these manual lists, however, are incomplete and efforts continue to find an algorithm to annotate words with sentiment information automatically (e.g., [38, 41, 60, 132]).

Automatic methods of sentiment annotation at the word level employ different techniques that can be grouped in two categories: (1) corpus-based approaches and (2) dictionary-based approaches. The first group includes methods that rely on syntactic or co-occurrence patterns of words in large texts to determine their sentiment (e.g., [50, 52, 62, 131, 156] and others). The reported accuracy of automatic systems for sentiment tagging of words is as high as 92%, which was reported for adjectives (the easiest part of English) with the exclusion of "difficult" words on which human annotators did not agree, while the lowest results reported on a broader set of words is as low as 62%. The evaluation methods also can play an important role in determining the reported system performance. The majority of dictionary-based approaches use WordNet relations, especially synsets and hierarchies, to acquire sentiment-marked words [56, 122], to create training sets for automatic sentiment classifiers [38], or to measure the similarity between candidate words and sentiment-bearing words such as *good* and *bad* [60].

This chapter introduces a new approach to sentiment extraction from WordNet at the word and the individual sense level by complementing the learning of unigram features using WordNet relations with learning from WordNet glosses[2] and lexical relations. The approach is applicable to the acquisition of sentiment-bearing words as well as of words of

[2] WordNet gloss is a definition of a synset meaning, for example, *good* in sense 1 is defined as *having desirable or positive qualities especially those suitable for a thing specified*; *good* in sense 2 (synset *full, good*) has gloss *having the normally expected amount.*

some other semantic categories, such as words with increase/decrease semantics and other valence shifters, which are also relevant for sentiment analysis. The resulting wordlists can then be used as an input for sentence and text-level sentiment analysis. This chapter, thus, first describes the WordNet-based approach to sentiment tagging at the word and sense level (Section 4.1), and then evaluates the obtained wordlist as part of the sentence and text-level sentiment-tagging system — alone and in combination with other information (Section 4.2).

4.1 Sentiment Tag Extraction from WordNet Glosses

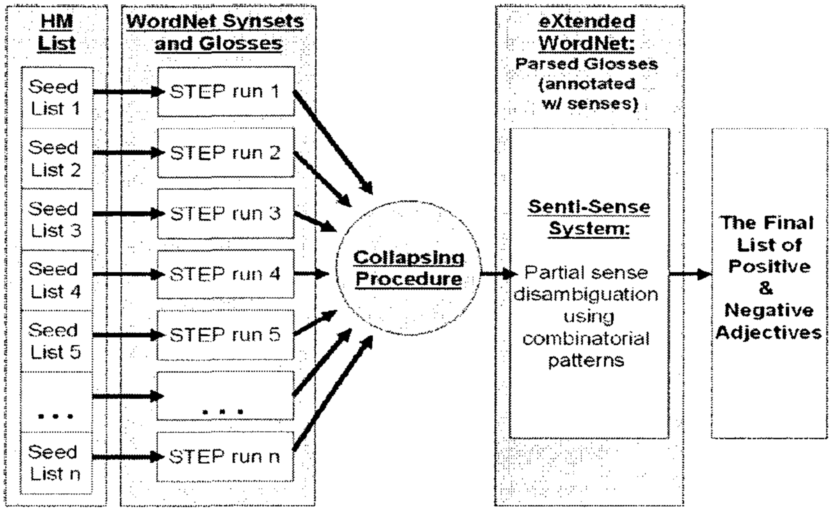

Figure 7: System architecture.

The approach to building a lexicon for sentiment analysis described here relies both on lexical relations (synonymy, antonymy and hyponymy) provided in WordNet and on WordNet glosses. It builds upon the properties of dictionary entries as a special kind of structured text: such lexicographical texts are built to establish semantic equivalence between the left-hand and the right-hand parts of the dictionary entry, and therefore are designed to match as close as possible the meaning components of the word. Their standardized style, grammar and syntactic structures remove a substantial source of noise common to other types of text. Finally, they have an extensive coverage spanning the entire lexicon of a natural language. The proposed approach to sentiment annotation of WordNet entries was implemented and tested in the Semantic Tag Extraction Program (STEP, Figure 7).

The STEP algorithm starts with a small set of seed words of known sentiment value (positive or negative). The list created by Hatzivassiloglou and McKeown [52] (HM) was used as a source of seed words for these experiments. This list is augmented during the first pass by adding synonyms, antonyms and hyponyms of the seed words supplied in WordNet. This step brings on average a 5-fold increase in the size of the original list with the accuracy of the resulting list comparable to manual annotations (78%, similar to HM vs. Harvard-IV list from the General Inquirer (GI-H4) accuracy, table 11 and 12). For example, a seed adjective *good#a*[3] with its 21 senses will fetch *full, estimable, honorable, respectable, beneficial, just, upright, adept, expert, practiced, proficient, skillful, skilful, dear, near, dependable, safe, secure, right, ripe, well, effective, serious, sound, beneficial, etc.* At the second pass, the system goes through all WordNet glosses and identifies the entries that contain in their definitions the sentiment-bearing words from the extended seed list and adds these head-words to the corresponding category — positive, negative or neutral (the remainder). For instance, *good* will bring, among others, ***lucky*** — *blessed with good fortune* and ***unhealthy*** — *not in or exhibiting good health in body or mind.* The approach also includes negation handling that reverses the sentiment of a seed word if this word follows a negation, as in the gloss of *unhealthy* above.

A third, clean-up pass is then performed to partially disambiguate the identified WordNet glosses with Brill's part-of-speech tagger [21], which performs with up to 95% accuracy, and eliminates errors introduced into the list by part-of-speech ambiguity of some words acquired in pass 1 and from the seed list. For instance, *sound* as a verb or noun is sentiment-neutral, while as an adjective it has a positive sentiment (e.g., *sound practice*). Thus, after this step the system will keep ***intelligent, well-informed*** — *possessing sound#a knowledge*, but will eliminate ***hearing*** — *able to perceive sound#n.*

At this step, the system also eliminates all words that have been assigned contradicting, (i.e., positive and negative) sentiment values, within the same run.

The performance of STEP was evaluated against the Harvard-IV list from the General Inquirer (GI-H4) as a gold standard. This list consisted of 1904 adjectives from GI-H4 and included not only the words that were marked as *Positiv, Negativ,* but also those that were not considered sentiment-laden by GI-H4 annotators. The latter, thus, were by default considered neutral in our evaluation.

4.1.1 Experiments

Two sets of experiments were conducted in order to assess the performance of the algorithm. In the first experiment, the algorithm was run with the full seed list (HM) over all of

[3] In this chapter I am using the WordNet notation of the following structure: word#part of speech#sense number.

WordNet (Table 11). The performance of the algorithm is 10% below the inter-annotator agreement for two manually annotated lists. Both human agreement and accuracy of the automatic system are considerably higher for the classification without neutrals, confirming that it is more difficult to separate neutrals from sentiment-bearing words than to distinguish between positive and negative sentiment.

List	STEP vs.		GI vs.
source	GI	HM	HM
Intersection size	623	188	744
Accuracy WITH neutrals	66%		79%
Accuracy WITHOUT neutrals	88%	91%[4]	98%

Table 11: Accuracy on the intersection: STEP vs. GI and HM.

In the second set of experiments, for the purposes of deeper analysis and evaluation, the entire HM list was randomly partitioned into 58 same-size non-intersecting seed lists of adjectives. The results of the 58 runs on these non-intersecting seed lists are presented in Table 12. Table 12 shows that the performance of the system exhibits substantial variability depending on the composition of the seed list, with accuracy ranging from 47.6% to 87.5% percent (Mean = 71.2%, Standard Deviation (St.Dev) = 11.0%).

The partitioning of the full list of HM adjectives into small sub-lists allowed us to obtain the results on a higher level of granularity than it was possible with the full list. Also, running these batches in parallel significantly sped up the computation.

	Average run size		Average % correct	
	#of adj	St.Dev.	%	St.Dev.
PASS 1 (WN Relations)	103	29	78.0%	10.5%
PASS 2 (WN Glosses)	630	377	64.5%	10.8%
PASS 3 (POS Clean-up)	435	291	71.2%	11.0%

Table 12: Performance statistics on STEP runs.

The significant variability in accuracy of the runs (Standard Deviation over 10%) is attributable to the variability in the properties of the seed list words in these runs. The

[4] HM does not contain neutrals therefore we can compare only accuracy of the tags on sentiment-bearing words. This number is below 91% because in some cases the sentiment coming from the seed list is overwritten by the system results, e.g. *unsuspecting* is labeled negative in HM, while in WordNet its meaning is either neutral (*not knowing or expecting; not thinking likely*) or slightly positive (*not suspicious*); *indebted* is positive in one sense (*owing gratitude or recognition to another for help or favors etc* and neutral in another (*under a legal obligation to someone*), while HM labeled it as negative.

HM list (the seed list in the experiments) includes some sentiment-marked words where not all meanings are laden with sentiment; words where some meanings are neutral; and even words where such neutral meanings are much more frequent than the sentiment-laden ones. The runs where seed lists included such ambiguous adjectives were labeling many neutral words as sentiment marked since such seed words were more likely to be found in the WordNet glosses in their more frequent neutral meaning. For example, run # 53 had in its seed list two ambiguous adjectives *dim* and *plush*, which are neutral in most of the contexts. This resulted in only 52.6% accuracy (18.6% below the average). Run # 48, on the other hand, by chance, had only unambiguous sentiment-bearing words in its seed list, and, thus, performed with a high accuracy (87.5%, 16.3% above the average).

In order to generate a comprehensive list covering the entire set of WordNet adjectives, the 58 runs were then collapsed into a single set of unique words. Since many of the clearly sentiment-laden adjectives that form the core of the category of sentiment were identified by STEP in multiple runs and had, therefore, multiple duplicates in the list that were counted as one entry in the combined list, the collapsing procedure resulted in a lower-accuracy (66.5% - when GI-H4 neutrals were included), but also in a much larger list of English adjectives marked as positive ($n = 3,908$) or negative ($n = 3,905$). The remainder of WordNet's $22,141$ adjectives was not found in any STEP run and hence was deemed neutral ($n = 14,328$) by default.

Table 11 shows the accuracy for the intersection between the two manually created word lists compared to each other. The agreement between human annotations in GI and HM was only 79% (when neutrals are taken into account), which suggests that at least for some semantic categories (such as sentiment) the rate of inter-annotator agreement cannot be expected to be high, unless the annotators are trained to code similar cases in a uniform way. This observation may have important implications for the evaluation of semantic tagging systems.

The errors on positive-neutral and negative-neutral boundaries accounted for the bulk of system errors and almost all of the disagreement between the two human-annotated lists. This suggests that the boundaries between the coding categories (positive vs. neutral vs. negative) are fuzzy and both humans and computer systems show much smaller rates of agreement on positive vs. neutral or negative vs. neutral distinctions than on positive vs. negative labels. This problem is addressed by Turney and Littman, [132], Grefenstette et al. [50] and Kamps et al. [60] by setting a scoring threshold below which words are deemed neutral.

The analysis of STEP system performance vs. GI-H4 and of the disagreements between manually annotated HM and GI-H4 showed that the greatest challenge with sentiment tagging of words lies at the boundary between sentiment-marked (positive or negative) and sentiment-neutral words. The 22% performance gain (from 66% to 88%) associated with the removal of neutrals from the evaluation set emphasizes the importance of neutral words as an important source of sentiment extraction errors. It is consistent with the observation by Kim and Hovy [62] who noticed that when positives and neutrals were collapsed into the same category (opposed to negatives), the agreement between human annotators rose by 12%. Moreover, the boundary between sentiment-bearing (positive or negative) and neutral words in GI-H4 accounts for 93% of disagreements between the labels assigned to adjectives in GI-H4 and HM by two independent teams of human annotators. The view taken here, however, is that the vast majority of such inter-annotator disagreements are not really errors but a reflection of the natural ambiguity of the words that are located on the periphery of the sentiment category. The inter-annotator studies, thus, should be regarded not just as a way to ensure accuracy of human annotations or set the upper bound for system evaluation, but also as an important method for identification of linguistic units (words, or n-grams) that are central/peripheral to a given semantic category.

4.1.2 Establishing the Degree of Word's Centrality to the Semantic Category

Sentiment as a Fuzzy Set

The category of sentiment, as many other linguistic variables, can be represented as a fuzzy set where words can have varying degrees of membership. Fuzzy logic has been introduced by Lotfi Zadeh [157] for modeling uncertainty of natural language. It extends the conventional Boolean logic in order to handle partial truth — the truth values that lie between "completely true“ and “compelely false”. Fuzzy is a pair (S, m) where S is a set and m is defined as $m : A \rightarrow [0, 1]$. For each $s \in S$, $m(s)$ is its degree of membership. If $m(s) = 0$ it means that x is not included in set S, $m(x) = 1$ means full membership in S. Values strictly between 0 and 1 characterize the fuzzy members. Use of fuzzy logic allows modeling different degrees of centrality (relatedness) of words to the semantic category of sentiment.

The proposed approach to the category of sentiment as a fuzzy set implies some additional structural properties of this category. First, as opposed to words located on the periphery, more central elements of the set usually have stronger and more numerous semantic relations with other category members[5]. Second, the membership of these central

[5] The operationalizations of centrality derived from the number of connections between elements can be found in social network theory [23]

words in the category is less ambiguous than the membership of more peripheral words. Thus, we estimate the centrality of a word in a given category in two ways:

1. Through the density of the word's relationships to other words — by enumerating its semantic ties to other words within the field, and calculating membership scores based on the number of these ties; and

2. Through the degree of word membership ambiguity — by assessing the inter-annotator agreement on the word membership in this category.

Lexicographical entries in the dictionaries, such as WordNet, seek to establish semantic equivalence between the word and its definition and provide a rich source of human-annotated relationships between the words. By using a bootstrapping system, such as STEP, that follows the links between the words in WordNet to find similar words, we can identify the paths connecting members of a given semantic category in the dictionary. Every run is done with a different seed list and, therefore, when a word is retrieved by multiple runs it means that it is related, through semantic ties or through gloss definitions, to several sentiment-laden words. This provides additional evidence in support of its membership in this category. With multiple bootstrapping runs on different seed lists, we can then produce a measure of the density of such ties. The ambiguity measure derived from inter-annotator disagreement can then be used to validate the results obtained from the density-based method of determining centrality: it can be expected that higher inter-annotator agreements corresponds to the higher degree of membership in the semantic category.

Net Overlap Score

In order to produce a centrality measure, I conducted multiple runs with non-intersecting seed lists drawn from HM. The lists of words fetched by STEP on different runs partially overlapped, suggesting that the words identified by the system many times as bearing positive or negative sentiment are more central to the respective categories. The number of times the word was retrieved by STEP runs is reflected in the *Gross Overlap* Measure produced by the system. In some cases, there was a disagreement between different runs on the sentiment assigned to the word. Such disagreements were addressed by computing the *Net Overlap* Scores (NOS) for each of the extracted words: the total number of runs assigning the word a negative sentiment was subtracted from the total of the runs that consider it positive. Thus, the greater the number of runs fetching the word (i.e., Gross Overlap) and the greater the agreement between these runs on the assigned sentiment, the higher the Net Overlap Score of this word. For example, for word *good* that is central to the category, the NOS is equal to 18 which means that in 18 runs *good* was retrieved as a word with a positive sentiment (it was part of the seed list for one of these runs and was extracted as a synonym of other seed words in 17 others). On the other hand, an *audacious* that is ambiguous in terms of sentiment (HM labeled it as positive, while GI considers it negative) has NOS=-1.

Relation between Net Overlap Score and Inter-annotator Agreement

The Net Overlap scores obtained for each identified word were then used to stratify these words into groups that reflect positive or negative distance of these words from the zero score. The zero score was assigned to (a) the WordNet adjectives that were not identified by STEP as bearing positive or negative sentiment[6] and to (b) the words with equal number of positive and negative hits on several STEP runs. The performance measures for each of the groups were then computed to allow the comparison of STEP and human annotator performance on the words from the core and from the periphery of the sentiment category. Thus, for each of the Net Overlap Score groups, both automatic (STEP) and manual (HM) sentiment annotations were compared to human-annotated GI-H4, which was used as a gold standard in this experiment.

On 58 runs, the system has identified 3,908 English adjectives as positive, 3,905 as negative, while the remainder (14,428) of WordNet's 22,141 adjectives was deemed neutral. Of these 14,328 adjectives that STEP runs deemed neutral, 884 were also found in GI-H4 and/or HM lists, which allowed evaluation of STEP performance and HM-GI agreement on this subset of neutrals as well. The graph in Figure 8 shows the distribution of adjectives by Net Overlap scores and the average accuracy/agreement rate for each group.

The graph in Figure 8 shows the relationship between NOS (x axis), size of the corresponding "bucket" (on the right) and sentiment annotation accuracy (on the left). The lighter line represents STEP accuracy against GI, the darker one — agreement between HM and GI on the same set of adjectives. The "buckets" group adjectives with the same NOS and accuracy and inter-annotator agreement were computed separately for each such group. The largest group — neutral adjectives — with $NOS = 0$ have the lowest accuracy both for automatic and manual annotation. As we move from NOS=0 to higher values of 0, the number of adjectives having this score diminishes while the accuracy and inter-annotator agreement increases.

From Figure 8 we can observe that the greater the Net Overlap Score, and hence,

[6] The seed lists fed into STEP contained positive or negative, but no neutral words, since HM, which was used as a source for these seed lists, does not include any neutrals.

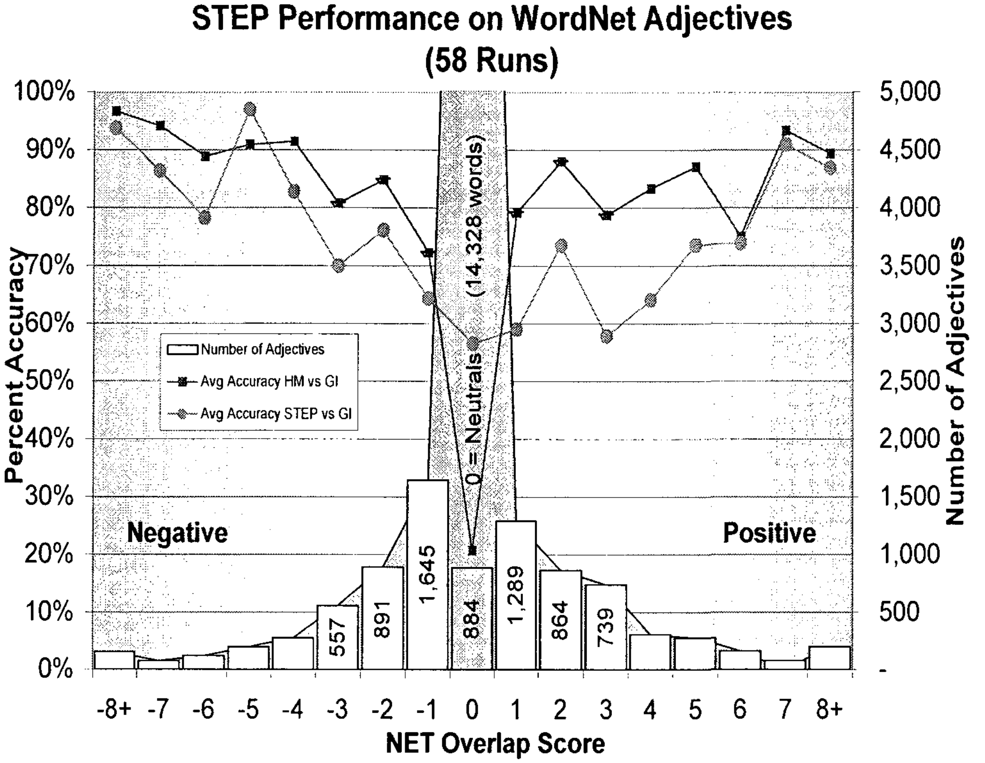

Figure 8: Accuracy of word sentiment tagging stratified by NOS values.

the greater the distance of the word from the neutral subcategory (i.e., from zero), the more accurate are STEP results and the greater is the agreement between two teams of human annotators (HM and GI-H4). On average, for all categories, including neutrals, the accuracy of STEP vs. GI-H4 was 66.5%; human-annotated HM had 78.7% accuracy vs. GI-H4. For the words with Net Overlap of ±7 and greater, both STEP and HM had accuracy around 90%. The accuracy declined dramatically as Net Overlap scores approached zero (= Neutrals). In this category, human-annotated HM showed only 20% agreement with GI-H4, while STEP, which deemed these words neutral, rather than positive or negative, performed with 57% accuracy.

These results suggest that the two measures of word centrality, Net Overlap Score (NOS) based on multiple STEP runs and the inter-annotator agreement (HM vs. GI-H4), are directly related[7]. Thus, the Net Overlap Score can serve as a useful tool in the identification of core and peripheral members of a fuzzy lexical category, as well as in prediction of inter-annotator agreement and system performance on a subgroup of words characterized by a given Net Overlap Score value.

Fuzzy membership in the category of sentiment

The absolute values of NOS may vary depending on the number of the separate experiments used. In order to make the Net Overlap Score measure more uniform and easier to use in sentiment tagging of texts and phrases, the absolute values of this score should be normalized and mapped onto a standard $[0, 1]$ interval. Since the values of the Net Overlap Score may vary depending on the number of runs used in the experiment, such mapping eliminates the variability in the score values introduced with changes in the number of runs performed. In order to accomplish this normalization, I used the value of the Net Overlap Score as a parameter in the standard fuzzy membership S-function [157, 158]. This function is commonly used for linguistic variables. It maps the absolute values of the Net Overlap Score onto the interval from 0 to 1, where 0 corresponds to the absence of membership in the category of sentiment (in our case, these will be the neutral words) and 1 reflects the highest degree of membership in this category. The S-function is defined as follows:

$$S(u;\alpha,\beta,\gamma) = \begin{cases} 0\,\mathrm{for}\, u \leq \alpha \\ 2(\frac{u-\alpha}{\gamma-\alpha})^2\,\mathrm{for}\, \alpha < u < \beta \\ 1 - 2(\frac{u-\gamma}{\gamma-\alpha})^2\,\mathrm{for}\, \beta \leq u < \gamma \\ 1\,\mathrm{for}\, u \geq \gamma \end{cases}$$

where u is the Net Overlap Score for the word and α, β, γ are the three adjustable parameters: α is set to 1, γ is set to 15 and β, which represents a crossover point, is defined as $\beta = (\gamma + \alpha)/2 = 8$. The values of α and β were chosen based on the observations made on the data and the purpose of the list: the words with NOS=1 are borderline and thus then between sentiment-laden and neutral and there was very little evidence that supports their inclusion in the category of sentiment, therefore these words should not be considered by a sentence- or text-level sentiment-tagging system. The value of β was set to 15 because there was no difference in the behavior of words with NOS$\geq$15. Defined this way, the S-function assigns highest degree of membership (=1) to words that have the Net Overlap Score $u \geq 15$. The accuracy vs. GI-H4 on this subset is 100%. The accuracy goes down as

[7] In our sample, the coefficient of correlation between absolute values of NOS and inter-annotator agreement was 0.68. The *Absolute* Net Overlap Score (not translated into degrees of fuzzy membership) on the subgroups 0 to 10 was used in calculation of the coefficient of correlation.

the degree of membership decreases and reaches 59% for values with the lowest degrees of membership (see Table 13).

Degree of membership S	Average accuracy
$S = 0$	58.9%
$0 < S < 0.5$	73.1%
$0.5 \leq S < 1$	86.6%
$S = 1$	100%

Table 13: Accuracy of word-level sentiment tagging for different degrees of membership in the fuzzy set of sentiment.

Since low degrees of membership are associated with greater ambiguity and inter-annotator disagreement, the Net Overlap Score value can provide researchers with a set of volume/accuracy trade-offs. For example, by including only the adjectives with the Net Overlap Score of 4 and more, the researcher can obtain a list of 1,828 positive and negative adjectives with accuracy of 81% vs. GI-H4, or 3,124 adjectives with 75% accuracy if the threshold is set at 3.

4.1.3 Sense-level Tagging

Another important reason for the low accuracy for words at the boundary between sentiment-bearing and neutral categories is that the same word can have both sentiment-bearing and neutral senses, and aggregating the tags at the word-level often leads to questionable tags and errors in sentence and text-level sentiment annotation. For example, there are a considerable number of adjectives that have both neutral and sentiment-laden meanings, and the neutral meaning is often a more frequent one. Table 14 shows frequency scores assigned to a subset of these adjectives in WordNet [42]. The scores reflect the number of occurrences of a given sense in semantic concordances created by WordNet editors.

Word	Total frequency of neutral senses	Total frequency of sentiment senses	Total occurrences[8]
great	141(75%)	46(25%)	187
dark	90(90%)	10(10%)	100
cold	41(76%)	13(24%)	54
right	20(28%)	50(71%)	70

Table 14: Frequencies of sentiment-marked and neutral meanings (based on WordNet).

Table 14 shows that in the corpus used by WordNet authors, on every occurrence of the word *great* that manual lists classify as positive [52, 118], there is a 75% chance that

[8] The number in this column is the sum of all WordNet frequency scores for all the senses of the given word. If there was no frequency assigned to the sense in WordNet it was considered to be equal to 0.

this word is used in one of its neutral, non-sentiment-bearing meanings. This could lead to sentiment annotation system error in up to 75% of the cases where *great* is found in a text. While 75% error rate attributable to the presence of neutral senses represents an extreme case, the error rate of 20% to 50%, according to our manual analysis, appears to be very common for polysemous words with at least one sentiment-laden meaning. Since the sentiment of each meaning for every English word has not been identified yet, this problem cannot be addressed by sentiment aggregation to the word level using probabilities of the occurrence of a given sense in a text: to date, the only way to arrive at the conclusion that a word appears in texts in neutral meanings at a certain rate is through manual annotation at the sense level. An additional problem with such aggregation is that it would still lead to a substantial number of errors where words used in sentiment-bearing meanings are deemed neutral.

Thus, the indiscriminate inclusion of such adjectives at the word level into the lists of sentiment-bearing words that are then used as sentiment markers in sentiment tagging of phrases and texts introduces errors and has a detrimental effect on the overall system performance. This problem is often further exacerbated by the high frequency of such words in natural language. The analysis of this example suggests that sense-level annotation of sentiment markers can substantially improve the accuracy of sentiment tagging of texts and phrases[9].

The analysis of STEP's output showed that most errors made by the system occur at the boundary between neutral and sentiment-marked adjectives, while the confusion between positive and negative sentiment was rare. In many cases, seed adjectives fed into STEP (e.g., *great, cold*, etc.) had both neutral and sentiment marked meanings and, therefore, could appear in the glosses of sentiment-bearing as well as neutral words, for example:

> *redheaded* — having red hair and unusually **fair** skin.
> *sporty* — exhibiting or calling for sportsmanship or **fair** play.

This led to erroneous inclusion of neutral adjectives into either the positive or negative categories. Therefore, the two critical tasks that the system had to address were (1) the improvement of differentiation between neutral and sentiment-bearing adjectives, and (2) the identification of occurrences of sentiment-laden meaning(s) of a given polysemous adjective from the occurrences of its neutral meaning(s). In order to address these tasks, I developed a filtering procedure that builds upon the observation that nouns are good indicators of the senses of adjectives that modify them [10, 154]. Consider the following example: adjective

[9] [61] use more fine-grained annotations for rare cases where the same word can have positive and negative meanings, but they are not considering words that can be both neutral and sentiment-bearing.

fair is sentiment-neutral when it is combined with concrete nouns (*fair skin*), but it is used in its sentiment-bearing sense when it is combined with abstract nouns like *fair game, fair dealing*, etc. This pattern is observed for a number of other adjectives (e.g., dark, mild, cold). Such repeated co-occurrence patterns can help to differentiate between word meanings. In this study, I used co-occurrence patterns of nouns and adjectives to differentiate neutral and sentiment-bearing adjectives at the meaning level.

A number of approaches successfully use syntactic patterns to assign semantic tags, such as 'subjectivity', 'humans', 'locations', 'buildings', etc., to words [106, 128]. [106] describe a machine learning approach used in the Meta-bootstrapping and Basilisk systems for learning subjective nouns based on an extraction pattern bootstrapping algorithm. In both systems, the algorithm starts with a seed list of subjective nouns and an unannotated corpus to create a pool of patterns which, in turn, is used to classify other nouns as 'subjective' or 'objective'. [52] developed a system that was using a specific co-occurrence pattern — a conjunction linking two adjectives — to draw conclusions about the sentiment value of one member of the pair based on the known sentiment of the other member. Automated pattern learning and matching has also been employed in general word-sense disambiguation systems (e.g., [83]).

Learning Generalized Co-occurrence Patterns

The approach described here (Figure 9) uses machine learning in order to generate and generalize the patterns that permit differentiation between sentiment-laden and neutral adjectives based on the semantic category of the noun they modify. The system that I developed for partial word sense disambiguation in sentiment-bearing words using combinatorial patterns (thereafter Senti-Sense system) is an extension of the STEP system described in the previous section. It proceeds as follows: First, the list of all non-ambiguous, monosemous sentiment-bearing adjectives is extracted from HM: all adjectives in HM that have only one sense in WordnNet were included in this list. The resulting list of 363 positive or negative adjectives is used as a seed list for the pattern extraction algorithm. As was the case in STEP, this seed list is then augmented by all synonyms of these words found in their WordNet synsets, which resulted in 1019 word-senses. All entries in this extended list have a WordNet sense number assigned to them[10].

The system takes advantage of the additional information contained in extended Wordnet. The eXtended WordNet [85] provides parses for most WordNet glosses as well as part-of-speech and sense assignments for words in them. The project is currently under

[10] It is also possible to use synset offset to identify the sense by the synset it belongs to, but I chose to record the sense numbers instead, because eXtended WordNet uses this notation.

1. Start with a seed list made of manually annotated non-ambiguous adjectives,
2. Expand it with synonyms and antonyms using WordNet relations;
3. Extract adjective-noun pairs for seed adjectives;
4. Get hypernyms of the nouns modified by the seed adjectives;
5. Generalize patterns by grouping the nouns by the lowest common hypernym.

Figure 9: Algorithm for deriving the patterns for sentiment adjective disambiguation.

development (the current version is XWN 2.0-1) and the quality of the tags varies from manual annotations ("gold" quality) to results of multiple automatic disambiguation systems ("silver" quality) to single system output ("normal" quality). Nevertheless, it still provides useful information that can be used to discover and apply syntactic patterns for sentiment tagging. In this research, disambiguated glosses serve as a corpus for learning and applying the co-occurrence patterns.

The system then searches the eXtended WordNet (XWN) for glosses that contain adjectives from this sense-tagged list. If an adjective is found, the parses and sense-tags provided in eXtended WordNet are used to identify the noun that this adjective modifies as well as the sense in which this noun is used in the text portion of the WordNet gloss. In addition, the full hypernym ancestry of that noun is extracted from WordNet's hypernym hierarchy.

Once all adjective-noun pairs have been extracted from XWN, the Senti-Sense system performs pattern induction and matching. It groups the nouns that can take on sentiment-bearing modifiers, identifies common hypernyms of these nouns, and then generalizes the adjective-noun pattern to the level of that hypernym. The example below illustrates the generalization algorithm on the example of several adjective-noun pairs.

Adjective-noun pair 1:

selfish $_{a1}$ **actor** $_{n1}$

Noun hypernym tree:

$actor_{n1} \rightarrow performer_{n1} \rightarrow entertainer_{n1} \rightarrow person_{n1} \rightarrow organism_{n1} \rightarrow living\_thing_{n1} \rightarrow object_{n1}$

Adjective-noun pair 2:

high-ranking $_{a1}$ **administrator** $_{n1}$

Noun hypernym tree:

$\mathit{administrator}_{n1} \rightarrow \mathit{head}_{n4} \rightarrow \mathit{leader}_{n1} \rightarrow \mathit{person}_{n1} \rightarrow \mathit{organism}_{n1} \rightarrow \mathit{living\_thing}_{n1} \rightarrow \mathit{object}_{n1}$

Adjective-noun pair 3:

skillful$_{a1}$ **opponent** $_{n2}$

Noun hypernym tree:

$\mathit{opponent}_{n2} \rightarrow \mathit{person}_{n1} \rightarrow \mathit{organism}_{n1} \rightarrow \mathit{living\_thing}_{n1} \rightarrow \mathit{object}_{n1}$

Common segment of hypernym tree for the three nouns:

$\rightarrow \mathit{person}_{n1} \rightarrow \mathit{organism}_{n1} \rightarrow \mathit{living\_thing}_{n1} \rightarrow \mathit{object}_{n1}$

The three adjective-noun pairs presented above were returned by the search for occurrences of the unambiguous sentiment-marked adjectives *selfish, high-ranking*, and *skillful.* For every occurrence of these adjectives in WordNet glosses, the system pulled out the adjective, the noun it modified, and the complete list of hypernyms in the WordNet hierarchy for this noun. The hypernym trees leading to each of the three nouns — *actor, administrator*, and *opponent* — converged at the hypernym *person* and reached the highest level hypernym *object*[11]. Thus, the system derived the pattern that nouns found under the hypernym *person* could take sentiment-bearing modifiers. Some of the patterns that were observed on the data were overly general. For instance, a very frequent pattern where sentiment adjective modifies a noun with a hypernym *living_thing.*

Based on the 1006 adjective-noun pairs found in XWN, the system produced 48 generalized patterns that were then used to evaluate the positive and negative adjectives retrieved by the multiple STEP runs[12]. For example, the adjective *bright* identified as positive by STEP when fed into the Senti-Sense system, returned a candidate pair bright_{a3} person_{n1}, where $\mathit{person}_{n1} \rightarrow \mathit{organism}_{n1} \rightarrow \mathit{living\_thing}_{n1} \rightarrow \mathit{object}_{n1}$. Given the pattern rule formulated in the example above, the Senti-Sense system concluded that the adjective *bright* in the sense 3 is sentiment-bearing. The WordNet gloss for *bright* in sense 3 is *characterized by quickness and ease in learning.* At the same time, another pair with the same adjective — bright_{a2} $\mathit{plumage}_{n1}$, where $\mathit{plumage}_{n1} \rightarrow \mathit{body\_covering}_{n1} \rightarrow \mathit{covering}_{n1} \rightarrow \mathit{natural\_object}_{n1} \rightarrow \mathit{object}_{n1}$, was not matched to any pattern characteristic of sentiment bearing words, and the adjective *bright* in sense 2 was deemed neutral. Table 15 provides examples of most frequent generalized patterns obtained from XWN.

[11] In order to avoid overgeneralization and to maintain the discriminating ability of the patterns, the top of the hypernym tree is not included in matching. Thus, in this case, no generalization to the highest hypernym level *object* was made.

[12] The generalization step has considerably reduced the number of observed patterns since multiple specific word pairs were replaced by a single common pattern.

Pattern	Frequency	Example
SENTIMENT_ADJ + psychological_feature#*n*#1	83	keen insight
SENTIMENT_ADJ + person#*n*#1	80	unpredictable voter
SENTIMENT_ADJ + social_relation#*n*#1	70	bad joke
SENTIMENT_ADJ + communication#*n*#2	69	guilty plea
SENTIMENT_ADJ + cognition#*n*#1	64	incorrect concept
SENTIMENT_ADJ + activity#*n*#1	59	unsportsmanlike foul
SENTIMENT_ADJ + state#*n*#4	52	deserving honor
SENTIMENT_ADJ + property#*n*#3	28	dangerous level
SENTIMENT_ADJ + social_group#*n*#1	25	efficient management

Table 15: Examples of generalized patterns observed on XWN. See Appendix 1 for the full list of generalized patterns.

Since positive and negative adjectives usually modify nouns from the same semantic classes, the Senti-Sense system can be effective in differentiating neutral and sentiment-laden adjectives, but not positives and negatives. Nevertheless, since the boundary between neutral and sentiment-bearing words represents the greatest source of errors in word sentiment determination, Senti-Sense can be a valuable extension for sentiment tag extraction systems, such as STEP. Moreover, given that Senti-Sense operates at the word meaning level, it can be used to identify which senses of a given adjective actually bear the sentiment value, and, by bringing sentiment tagging to the finer-grained sense level, it can further contribute to the accuracy of sentiment tagging systems.

Table 16 summarizes the results produced by these STEP runs before and after cleaning with Senti-Sense. I evaluated the tags produced by the multiple STEP runs against the General Inquirer Harvard IV list of 1904 adjectives. The GI list also contains adjectives that were not classified as "Positiv" or "Negativ", and, by default were considered in our evaluation as neutrals. The system performance was evaluated only on the words that are present both in our results and in GI. Since Senti-Sense assigns the sentiment tags at the sense level and the GI list is annotated with sentiment at the word level, for the purposes of this evaluation only, the sense-level positive or negative tags were reassigned to the word: if at least one sense of an adjective was classified by Senti-Sense as sentiment-laden, the whole word was deemed to have that sentiment value. The words that were left in the neutral category, thus, had only neutral senses.

Since there is a substantial inter-annotator disagreement in word sentiment annotation, the evaluation of machine-made annotations should take into account the baseline level of inter-annotator agreement that can be achieved in this task by independent teams of human annotators. The agreement statistics between the manually annotated lists of adjectives created by Hatzivassiloglou and McKeown and the General Inquirer team of annotators is presented in the table below for comparison.

	Baseline: STEP vs. GI	Senti-Sense vs. GI	Gold: HM vs. GI
Total tagged words	1904	1415	774
Agreement on tags	1267	1081	609
% of same tags	66.5%	76.6%	78.7%

Table 16: Performance of Senti-Sense approach compared to the baseline and the gold standard.

Overall, the adjective list obtained after the cleaning procedure was smaller, but its precision increased considerably: differentiating between sentiment-marked and neutral senses of seed words using the generalized adjective-noun patterns added another 10% to the accuracy on the intersection between the system results and the gold standard (Table 16). Since Senti-Sense was based on NP patterns that identify only sentiment-laden senses, filtering was done only on those adjectives that were classified by STEP as positive or negative, while the composition of the category of neutrals was left intact. Table 17 provides detailed statistics for each of the three categories: positive, negative, and neutrals. Lower recall observed in Senti-Sense results relative to STEP is explained by the fact that adjectives that were marked as sentiment-bearing by STEP but for which no occurrences in NP-patterns were found were not included in any of the three groups, since no definite conclusion about their sentiment could be made on this data. The exclusion of those adjectives from the reported results allows to see the incremental improvement in system performance on those sentences, where pattern geniralization would apply.

Sentiment	Baseline: STEP vs. GI			Senti-Sense vs. GI			Gold: HM vs. GI		
	P	R	F	P	R	F	P	R	F
Positive	0.68	0.77	0.72	0.95	0.61	0.74	0.90	0.51	0.67
Negative	0.75	0.70	0.72	0.91	0.54	0.68	0.98	0.46	0.63
Neutral	0.57	0.53	0.55	0.57	0.53	0.55	n/a[13]	n/a	n/a

Table 17: Precision (P), recall (R) and F-score (F) on three sentiment categories.

The absence of an exhaustive list of sentiment-tagged words (which makes the task of machine-made annotation meaningful) does not allow the reliable assessment of the recall measure on STEP and Senti-Sense outputs. The closest available proxy is the GI list itself: the proportion of GI adjectives correctly identified by the system as positive, negative or neutral and the total number of GI adjectives that were not found can give an idea of

[13] HM does not contain neutrals.

system performance on this measure. System precision and recall before and after Senti-Sense filtering are presented in Table 17. Table 17 shows that the gain associated with Senti-Sense filtering of positive and negative adjectives was substantial: the precision on positive adjectives increased from 68% to 95%, while on negatives it went up from 75% to 91%. These results are comparable to the precision of human annotation (Table 17). This gain, however, came with a considerable reduction in the size of the filtered list: Senti-Sense filtering reduced the list of sentiment-laden adjectives found by multiple STEP runs from 7813 to 2907. The gain in precision accompanied by the drop in recall by 16% for both positives and negatives, has left the F-scores practically unchanged: it went up 2% for positives, where the gain in precision was particularly large (27%), and down 4% for negatives, where the increase in precision was equal to the loss in recall.

4.1.4 Evaluation

The evaluation of a word-level sentiment tagging method presents a number of interesting challenges. First of all, there is no clear baseline performance that can be used for comparison. The agreement between human annotators can serve as an upper boundary, however, such comparison is not fully indicative of the automatic system performance. The major limitation of such comparison consists in the small size and non-random composition of existing manual lists. Due to the high cost of manual annotation, the lists of sentiment bearing words that can be used as a gold standard are relatively small. At the same time, these lists cannot be considered a random sample of the category since the words for annotation were selected based on some theoretical or practical consideration such as their frequency in a certain corpus. Therefore, the agreement between two such lists may depend on a variety of factors such as the corpus used by the annotators. Moreover, the coverage of manually annotated lists is small and different lists overlap only in a small portion of the vocabulary they include. According to the study conducted by Grefenstette et al. [50], only 20-25% of manual lists overlap. Our own observation on GI and HM confirms this conclusion: less than half of adjectives found in HM were also part of the GI list and vice versa. Due to this, it is impossible to use the manually annotated lists to meaningfully evaluate the recall or coverage of automatic word labeling, thus recall and F-measure cannot be computed for such lists. Thus, precision and accuracy are the only measures that can be computed for this data. Most work in word-level sentiment tagging evaluates the performance of automatic sentiment tagging methods at word level in terms of accuracy and therefore it will be used here as well. Following the established tradition in sentiment analysis research, the General Inquirer lexicon is used here as a gold standard for evaluation. GI has two important properties that set it apart from other manually annotated lists: its coverage and presence of words that have no sentiment. The latter quality makes this list the only available resource for evaluation of the role of neutrals in sentiment tagging at the word level.

Another important consideration that influences the evaluation of sentiment tagging approaches at the word level is related to the fuzzy character of the category of sentiment: both for manual and for high quality automatic annotation the accuracy can be expected to be higher on words that belong to the core of the category and lower for those words that are close to the boundary. Our experiments with stratification of the category based on the Net Overlap Score has demonstrated that the agreement is higher on words that have high scores and lower when NOS approaches 0 (Figure 8).

Finally, the evaluation of sense-level tags is not yet possible since there are no resources manually annotated for sentiment at the sense/synset level. Few attempts were made to evaluate sense-level sentiment tags [58, 41] against gold standards created specifically for the purposes of such evaluation, but in both cases the manually annotated resource was created post-factum, after the results of automatic tagging were already available, and was tailored to the specific approach used in the automatic tagging[14]. Therefore, the related resources are not suitable as gold standard for other similar applications. This lack of manually annotated resources contributes to the relevance of automatic sentiment tagging at sense level, but at the same time makes the evaluation and comparison of different approaches more difficult. Therefore, for the purposes of the evaluation, the sense-level sentiment was aggregated to the word level using the following procedure: the overall sentiment of a given word was computed as a sum of the sentiment of its senses, where each positive sense was counted as +1 and each negative sense as -1. If the resulting sum was greater than 0, the word was considered positive, if it was below 0 the word sentiment was deemed to be negative. Words with combined sentiment equal 0 were assumed to be neutral[15].

Overall, the STEP system's 88% accuracy in binary, positive vs. negative, classification of adjectives is superior to the accuracy reported in the literature for other systems run on large corpora [131, 52]. For instance, Turney and Littman [131] report 76.06% accuracy for experimental runs on 3,596 sentiment-marked GI words from different parts of speech using a 2x10^9 corpus to compute point-wise mutual information between the GI words and 14

[14] For instance, Esuli and Sebastiani [41] used a resource where sentiment was assigned on a scale from 0 to 1, which can not be directly translated into binary positive/negative values nor into values corresponding to NOS.

[15] That is, the neutrals at the word level included both "real" neutrality, when none of the word senses is sentiment-bearing, and ambiguity in respect of sentiment, when the same word has both positive and negative senses (e.g., *apt* that has both positive (*being of striking appropriateness and pertinence*) and negative (*at risk of or subject to experiencing something usually unpleasant*) senses).

manually selected positive and negative paradigm words. The accuracy of the approach on ternary classification of adjectives is better than the numbers reported by Kim and Hovy [66] for this category: average STEP accuracy in the 58 runs was 71.2% (Table 12), while Kim and Hovy [66] report accuracy of 69.1%. These numbers are, however, not directly comparable due to the differences in the gold standard.

The results are also significantly better (for binary, positive vs. negative, classification) or similar (for ternary classification) to those obtained by Esuli and Sebastiani [39]. In their approach, independently developed at the same time as this work, words in WordNet were classified into positive and negative based on their synsets, glosses and examples. Esuli and Sebastiani [39] used a number of statistical classifiers trained on the large number of WordNet entries. They report experiments with two training sets: the first consisted of 14 adjectives from [130], the second of adjectives *good* and *bad*; both training sets were extended by adding synonyms on multiple iterations. Esuli and Sebastiani [38] tested several classifiers: SVM, PrTFIDF (probabilistic version of Rocchio learner), and Naïve Bayes. The best results on GI's positive and negative words (neutrals excluded) as a gold standard were produced by PrTFIDF classifier (83%). Their best result for ternary classification (66%) was achieved by using two PrTFIDF classifiers: one to separate positive from non-positive words, and the other to differentiate negatives and non-negatives. Esuli and Sebastiani [40] assign a series of scores to each word: one value for each of three categories. This means, that every word is characterized by the probability of belonging to positive, negative, and neutral categories, e.g., **bad**(10) *capable of harming* has P=0.625, N=0.125, O=0.25. As opposed to NOS, such scores are hard to interpret and they do not reflect any real property of the word[16]. Recently, Esuli and Sebastiani [41] have also applied the PageRanking method to the classification of WordNet synsets by sentiment. The approach did not improve on the results of their earlier experiments.

4.1.5 Conclusions

The proposed approach to word- and sense-level sentiment annotation contributes to the development of NLP and semantic tagging systems in several respects.

- ***The structure of the semantic category of sentiment.*** The analysis of the category of sentiment of English adjectives presented here suggests that this category

[16] The authors offer three possible interpretations of these scores as a pure speculative exercise, without giving preference for any of them nor providing any theoretical foundation of assignment of these scores apart from operational one — since they use a probabilistic classifier that had to assign probabilities to each sense as part of the classification process. It is particularly hard to reconcile positive and negative probabilities assigned to the same sense as in **bad**(1) *below average in quality or performance* with P=0.375 N=0.25 and O=0.375

is structured as a fuzzy set: the distance from the core of the category, as measured by Net Overlap Scores derived from multiple STEP runs, is shown to affect both the level of inter-annotator agreement and the system performance vs. human-annotated gold standard.

- ***The list of sentiment-bearing words***. The list produced using the STEP algorithm spans the entire WordNet. The accuracy of the positive vs. negative classification at the word-level is superior to all reported approaches. The accuracy of ternary — positive vs. negative vs. neutral — classification is better or comparable to results produced by other algorithms.

  Since adjectives are the most relevant sentiment marker [53], the generation of the list of sentiment-laden adjectives was our special concern. For this part of speech, the list was cross-validated by multiple STEP runs containing 7,814 positive and negative English adjectives, with an average accuracy of 66.5%, while the human-annotated list HM performed at 78.7% accuracy vs. the gold standard (GI-H4)[17]. The remaining 14,328 adjectives were not identified as sentiment marked and therefore were considered neutral.

  Stratification of adjectives by their Net Overlap Score is an indicator of their degree of membership in the category of (positive/negative) sentiment. The normalization of the Net Overlap Score values for the use in phrase and text-level sentiment tagging systems was achieved using the fuzzy membership function that I proposed here for the category of sentiment of English adjectives.

- ***System evaluation considerations***. The contribution of this research to the development of methodology of system evaluation is two-fold.

  - First, this research emphasizes the importance of multiple runs on different seed lists for a more accurate evaluation of sentiment tag extraction system performance. I have shown how significantly the system results vary, depending on the composition of the seed list.

  - Second, due to the high cost of manual annotation and other practical considerations, most bootstrapping and other NLP systems are evaluated on relatively small manually annotated gold standards developed for a given semantic category. The implied assumption is that such a gold standard represents a random sample drawn from the population of all category members and hence, system

[17] GI-H4 contains 1268 and HM list has 1336 positive and negative adjectives. The accuracy figures reported here include the errors produced at the boundary with neutrals.

performance observed on this gold standard can be projected to the whole semantic category. Such extrapolation is not justified if the category is structured as a lexical field with fuzzy boundaries: in this case the precision of both machine and human annotation is expected to fall when more peripheral members of the category are processed. The sentiment-bearing words identified by the system were stratified based on their Net Overlap Score and evaluated in terms of accuracy of sentiment annotation within each stratum. These strata, derived from Net Overlap Scores, reflect the degree of centrality of a given word to the semantic category, and, thus, provide greater assurance that system performance on other words with the same Net Overlap Score will be similar to the performance observed on the intersection of system results with the gold standard.

- ***The role of the inter-annotator disagreement***. The results of the study presented in this dissertation call for reconsideration of the role of inter-annotator disagreement in the development of lists of words manually annotated with semantic tags. It has been shown here that the inter-annotator agreement tends to fall as we proceed from the core of a fuzzy semantic category to its periphery. Therefore, the disagreement between the annotators does not necessarily reflect a quality problem in human annotation, but rather a structural property of the semantic category. This suggests that inter-annotator disagreement rates can serve as an important source of empirical information about the structural properties of the semantic category and can help define and validate fuzzy sets of semantic category members for a number of NLP tasks and applications.

- ***Annotation at the sense-level***. The availability of a list of words annotated with sentiment tags at the sense, rather than word level, and the ability to use the Senti-Sense system to partially disambiguate adjectives in texts based on the semantic category of the noun they modify, opens up the possibility to develop more accurate sentiment tagging systems. Moreover, sentiment tagging systems that make use of sense-level sentiment information would be able to perform accurate tagging of small snippets of text (such as e-mails), where scarcity of lexical markers would hinder the effectiveness of sentiment tagging systems that rely on probabilistic assessment of multiple low-accuracy textual markers. The importance of the sense-level annotation has been recently recognized by several researchers [58, 142] and several approaches to sense-level semantic tagging were proposed [58, 142, 41]. However, most research in this direction is still at the exploratory stage and does not cover sufficient number of senses. The Senti-Sense algorithm was the first method for accurate assignment systematic of sentiment to a large number of synsets.

4.2 Lexicon-Based Approach to Sentence-level Sentiment Tagging

Comparison of the list of sentiment-bearing words to manually annotated resources is only the first step in the evaluation of such a wordlist. Since the purpose of the list is to provide features for sentiment classification of sentences and texts, the evaluation would have been incomplete without testing the list performance in this task. The resulting list of adjectives annotated with sentiment and with the degree of word membership in the category (as measured by the Net Overlap Score) will be used in sentiment tagging of phrases and texts. This will enable us to compute the degree of importance of sentiment markers found in phrases and texts. The availability of the information on the degree of centrality of words to the category of sentiment will improve the performance of sentiment determination systems built to identify the sentiment of entire phrases or texts.

The list of the words obtained using the STEP algorithm has been used in sentence and text level sentiment tagging of news and other domains (movie reviews, blogs, product reviews) in a series of experiments:

- **Baseline lexicon-based approach**: the sentiment of text, sentence or headline was computed as a difference in number of positive and negative words encountered in it.

- **Weighted lexicon-based approach**: instead of counting the number of positive and negative words, in this approach the system makes the decision about sentence sentiment based on Net Overlap Scores: it computes the difference between the sum of positive scores and the sum of negative scores for each sentence.

- **Syntax-aware lexicon-based system**: the weighted lexicon-based approach with an added syntactical component (dependency parse tree) and a list of valence shifters.

For the study of the lexicon-based approach (LBA) at sentence level, the list produced by STEP was cleaned following the procedure outlined in [106] who created lists of subjective nouns using bootstrapping and then manually reviewed the resulting list before it was used in sentence-level subjectivity classifier. This procedure was applied to the list produced by STEP in order to avoid compounding system errors at the sentence level with the errors produced at word level.

4.2.1 Baseline Lexicon-based Approach

The baseline lexicon-based approach (LBA) follows the procedure used by [63, 132, 56] and others. This method deduces sentiment of a sentence or text based on presence of sentiment-bearing words. The purpose of the system that uses this method was two-fold: first, it allowed the evaluation of the baseline lexicon-based approach, and second, it permitted exploration of the contribution of valence shifters handling to the overall system performance. Thus, it provided the baseline against which the contribution of other components such as valence shifter handling, was compared.

	Ternary classification	Binary classification			
	Accuracy PNO	Acccuracy PN	P	R	F
News	51.3	57.1	72.5	78.8	75.5
Blogs	56.3	59.5	78.5	75.8	77.1
Movie reviews	n/a	54.8	67.2	81.4	73.7
PRs	n/a	50.0	71.2	70.3	70.7

Table 18: Baseline LBA performance on sentences by genre (ternary (PNO) and binary (PN) classification.

Table 18 demonstrates one of important properties of a lexicon-based approach — its consistent performance across different domains and genres.

The performance of this system on news is lower than the performance of the approach used by Kim and Hovy [62]. They reported 67% accuracy on 100 sentences selected from DUC-01 Data. However, the results are not directly comparable since the approach used by Kim and Hovy [62] involved also the opinion holder identification and it is unclear what was the contribution of this additional feature to the overall performance of sentiment classifier. The system in [62] also used weights assigned to words based on their WordNet relations, while this baseline experiment presented here has not used the important contribution of the STEP algorithm — word ranking based on the Net Overlap Scores.

4.2.2 Weighted Lexicon-based Approach

The next series of experiments explores the role of Net Overlap Scores (NOS) in sentence-level sentiment tagging (Table 19[18]). The scores have significant impact on the system accuracy: the addition of scores leads to higher binary accuracy and recall and the corresponding increase in F-measure. The analysis of the results suggests that the main contribution of the scoring consists in breaking the ties when same number of positive and negative words

[18] The impact of the NOS-based weights is statistically significant for binary accuracy at $\alpha = 0.01$ for News and Product Reviews, at $\alpha = 0.025$ for Blogs, and 0.05 for Movie reviews.

is present in the sentence. The weighted lexicon-based approach relies on several sources of information. First of all, it draws on the results of the approach described in Section 4.1: the list of words (unigrams) generated using this approach constitutes is used as sentiment clues. The fuzzy Net Overlap Scores assigned to them serve to make these clues more accurate and to break the ties when the number of positive and negative clues in a sentence is the same. In this case, the sentiment determination at sentence and text level was done by summing up the scores of all identified positive unigrams (NOS scores > 0) and all negative unigrams (NOS scores < 0).

	Ternary classification	Binary classification			
	Acc PNO	Acc PN	P	R	F
News					
baseline	51.3	57.1	72.50	78.8	75.5%
weighted	51.1	61.9	71.9	86.1	78.4%
Blogs					
baseline	56.3	59.5	78.5	75.8	77.1%
weighted	58.4	63.5	78.6	80.8	79.7%
Movie reviews					
baseline	n/a	54.8	67.2	81.4	73.6%
weighted	n/a	57.3	64.9	88.4	74.8%
Product reviews					
Baseline	n/a	50.0	71.2	70.3	70.7
weighted	n/a	57.4	72.8	78.8	75.7

Table 19: Weighted Lexicon-Based Approach performance on sentences by genre.

Adding NOS scores to the words improves the accuracy and recall of LBA since more sentences are now annotated as non-neutral.

4.2.3 Lexicon-based Approach with Valence Shifters

Some language domains, such as news, abound in expressions that reverse or neutralize the sentiment conveyed by other words in the sentence, as in Sentence 9 in the text in Figure 1: *Although two candle fires were reported, no one was injured and no crime spikes occurred following the blackout, the mayor reported..* Here valence shifters[19] *although* and *no* neutralize the negative sentiment conveyed by other words in the sentence. The fuzzy Net Overlap Score counts were complemented with the capability to discern and take into account some relevant elements of syntactic structure of sentences. Two components were added to the system to enable this capability: (1) valence shifter handling rules and (2) parse tree analysis.

[19] See next section for the formal definition of valence shifters

Valence Shifter Handling

Valence shifters[20] can be defined as markers that modify the sentiment expressed by a sentiment-bearing word. They include negatives, intensifiers, modals, presuppositional items, irony and a number of discourse based elements [99]. The previous attempts to incorporate valence shifters in automatic systems produced mixed results: Kennedy and Inkpen [61] observed some improvement in system accuracy when valence shifters (negations and intensifiers) were taken into account, Wilson et al. [151] did not find noticeable difference in precision and recall of a classifier that had valence shifters included in the feature set and the one without them, and Dave et al. [32] reported negative impact of the inclusion of negation into the feature set. In contrast to the previous work that focused mainly on negations, I seek to create a list of valence shifters of different types and to study their interaction not only with the closest neighbor, but with subjective elements in a larger context.

The list of valence shifters for our experiments was compiled from three sources:

- a list of common English negations (such as *no, never, not*) taken from class 536 in [107],

- a subset of the list of automatically obtained words with increase/decrease semantics. The increase/decrease words are an important class of valence shifters - they modify the sentiment of other words in the sentence, by amplifying it (e.g., *their disappointment increased*) or diminishing it (e.g., *reduce the crime rate*). The words with increase/decrease semantics were acquired from WordNet using the same STEP algorithm as the one used to learn sentiment-bearing words.

- manual annotation by two annotators based on semi-automatic error analysis: when the baseline lexicon-based system was run on the development set, all sentences where a change of sentiment to the opposite was observed within the same sentence, were retrieved and analyzed by two annotators. For instance, in the sentence[21] *Today* $< 0 >$ *Kember* $< 0 >$ *responded* $< 0 >$ *to* $< 0 >$ *earlier* $< 0 >$ *criticism* $< -1 >$ *that* $< 0 >$ *they* $< 0 >$ *failed* $< -5 >$ *to* $< 0 >$ *show* $< 0 >$ *gratitude* $< 5 >$ *to* $< 0 >$ *their* $< 0 >$ *rescuers* $< 0 >$., two negative sentiment markers, *criticism* and *failed*, are followed by a positive word *gratitude*. The change of sentiment within the same sentence was considered a sign of possible presence of a valence shifter (in this case,

[20] The term "valence shifters" is the most widely used for this category of words and expressions, occasionally they are also called *polarity modifiers* and *polarity shifters* [151].

[21] Numbers in the angle brackets correspond to the NOS scored assigned to each word.

fail) and such sentences were reviewed by annotators. This allowed identifying several frequent valence shifters that were added to the list.

The full list consisted of 450 words and expressions.

Based on the way they modify the sentiment of other words in the sentence, I have identified four functional types of valence shifters based on their effect on sentiment of sentence constituents found within their scope: *valence reversers, valence neutralizers, valence painters, and intensifiers.*

Valence Reversers are words and expressions that reverse the sentiment of the constituents within their scope. Their scope, which usually (but not always) extends to the right of the valence shifter, can range from all words in the clause to a single verb, NP or AdjP phrase. E.g.,

> *The Board ordered that respondents* ***cease*** *each of the specified unfair labor practices.* (Positive: negative sentiment of *unfair* is reversed by the valence reverser *cease*) *Unions act* ***against*** *violence at work.* (Positive: the negative sentiment of *violence* is reversed as a result of the action of valence shifter *against*)

Valence Neutralizers can be regarded as a type of valence shifters that "turn off" the sentiment expressed by some other constituents of the sentence, for example, in

> *The weather on our trip was bad,* ***but*** *we enjoyed the company.* (Positive)

the sentiment of the sentence is dominated only by the part that follows *but.* In case of text-level sentiment analysis, *but* at the beginning of the sentence indicates that the sentiment of the previous sentence should be ignored.

Valence "painters" are sentiment-bearing words that "enforce" their sentiment to all words within their scope. They act in a way similar to neutralizers; instead of setting the sentiment of constituents within their scope to neutral, however, they "paint" even neutral words with some sentiment, e.g.,

> ***Unfortunately,*** *Richard Hall is correct on both counts.* (Negative)

Finally, **Intensifiers** are words that increase the strength of the expressed sentiment. Adverbs, such as *very, extremely,* and words with the semantics of "increase" belong to this category. Presence of intensifiers in a sentence leads to system errors for two reasons. First, there is no reliable algorithm that would measure the strength of each intensifier and represent it numerically. Second, the role of intensifiers can be clearly understood only when the system is using a perfectly clean, manually annotated word-list. Without a proper account of these two factors, the use of intensifiers may increase the impact of occasional errors in the lexicon on the overall system performance. Therefore, the words of this group were ignored in the experiments reported in this chapter.

For each of these functional categories of valence shifters (except intensifiers) I developed special handling rules that enabled our system to identify such words and phrases in the text and take them into account in sentence sentiment determination. Each word or expression in the list of valence shifters had their role and scope assigned to them in a following format: *Valence_shifter@action_to_take@node_where_the_scope_starts@scope*. For instance: *diminish@reverse_after@diminish@np* (as in *This action diminishes the vulnerability to fraud.*, where the negative sentiment of the entire NP governed by the verb *diminish* is reversed); *encouragingthat@pos_after@that@all* (as in *It is encouraging that Tim Geithner will be the new treasury secretary.*, where the entire, otherwise objective sentence is interpreted as positive because of the presence of *it is encouraging*). The system then reads in this information and whenever a valence shifter is encountered in a sentence, the corresponding action is applied to the specified part of the parse tree starting from the node indicated in the "trigger" field of the valence shifter entry. When several valence shifters are present in the sentence their actions are applied in a bottom-up fashion, for example, in the sentence *we did not waste a good day by settling at basecamp*, first the system applies valence shifter *waste* to NP *a good day* resulting in the negative sentiment of this part of the sentence, but then *not* is applied to the verb phrase, reversing its sentiment again and the resulting overall sentiment becomes positive.

The Table 20[22] shows that the introduction of valence shifter handling rules into the dictionary-based system has different impact on the classifier performance depending on the genre: The greatest gains were observed on the news, since news texts are particularly abundant in valence shifters. It is not uncommon to find news sentences with two or more valence shifters that interact with each other: e.g., *He is not without talent.* Movie and product reviews both gained in precision, but the recall diminished, resulting in a small decline in accuracy and F-measure. For blogs, the introduction of valence shifter handling had a negative impact on all measures. This negative effect of valence shifters handling on the blogs classification can in part be due to the complex and colloquial syntactic structures used by some blogs authors. For instance, it is hard to apply valence shifters correctly in such sentences as *It's not like it hurts that much anyways* or As far as good news goes, there is none, for now..

The addition of the valence shifters did not improve the accuracy of classification of

[22] The changes in accuracy compared to the weighted LBA were statistically significant only for Blogs ($\alpha = 0.025$). However, the difference between LBA with valence shifters and the baseline were statistically significant for all corpora ($\alpha = 0.01$ for News and PRs and 0.05 for Movie reviews), except Blogs.

	Ternary classification	Binary classification			
	Acc PNO	Acc PN	P	R	F
News					
baseline	51.3	57.1	72.5	78.8	75.5
weighted	51.1	61.9	71.9	86.1	78.4
weighted, valence shifters, no parse	54.3	62.8	76.4	82.2	79.2
Blogs					
baseline	56.3	59.5	78.5	75.8	77.1
weighted	58.4	63.5	78.6	80.8	79.7
weighted, valence shifters, no parse	56.9	59.5	75.8	78.5	77.1
Movie reviews					
baseline	n/a	54.8	67.2	81.4	73.6
weighted	n/a	57.3	64.9	88.4	75.3
weighted, valence shifters, no parse	n/a	57.5	67.2	85.5	75.3
Product reviews					
baseline	n/a	50.0	71.2	70.3	70.7
weighted	n/a	57.4	72.8	78.8	75.7
weighted, valence shifters, no parse	n/a	56.1	75.0	74.8	74.9

Table 20: Performance of LBA with valence shifters on sentences by domain and genre.

movie reviews and product reviews sentences. There were no comparable experiments at the sentence level in the extant literature, but Kennedy and Inkpen [61] studied the role of valence shifters in classification of movie review and product review texts. In their experiments, they augmented a set of sentiment clues with two kinds of valence shifters: negations and intensifiers. For movie reviews, the addition of these valence shifters resulted in 1.5-3% increase in accuracy, depending on the lexicon used, while for product reviews, it brought 0.5-2% improvement. In our experiments, however, the accuracy improved only on news sentences, while the accuracy on product and movie reviews did not change. This is probably, because on movie and product reviews, intensifiers play a greater role than other categories of valence shifters, and in our experiments, intensifiers were not used.

Parse tree

The use of parse tree information in extant literature on sentiment and subjectivity analysis is limited to specific constituents such *adjectival* or *complex noun phrases*. Thus, Bethard et al. [14] included only the presence of complex adjectival phrases as one of the features indicative of text subjectivity. Wilson et al. [151] include some "structure features" that reflect the element of the dependency tree to which the subjective element belonged (e.g., subject, copula, or passive). They also have a number of modification features that reflect only immediate modifiers of a word (e.g., is the word preceded by an adjective, an adverb, or an intensifier, or is it modified by a subjective word). Unlike these approaches, the method described here uses the dependency tree produced by MiniPpar [74] to determine the scope of valence shifters and to define phrase-level sentiment. It partially overlaps with the modification features in [151], but is more general since it covers all sentence elements and all relationships that may include sentiment markers and valence shifters, and thus is not limited to few special types of modifiers.

The actions that are associated with most valence shifters and some sentiment markers should be applied only to a specific constituent of the sentence, such as a direct object, adjectival phrase, etc. For instance, *not* modifies the sentiment of the phrase that follows the *not* (e.g., *"Day" is* ***not*** *a great bond movie*). There are valence shifters that work "backwards", changing the sentiment of the preceding phrase (e.g., *but* makes the preceding clause neutral as in *it's not exactly a gourmet meal,* ***but*** *the fare is fair*).

In order to correctly determine the scope of valence shifters in a sentence, I introduced into the system parse tree analysis using MiniPar [74] — a dependency parser that performs with 89% precision. The parser also assigns part of speech tags to words, which improves precision of sentiment-bearing unigram identification by partially disambiguating homonymous unigrams: It is not uncommon that in a pair of homonyms one homonym is sentiment-laden, while the other one belonging to a different part of speech is neutral. For example, *sound* has no sentiment as a noun or verb but is positive as an adjective. The ability of the system to discern between *sound* as a noun and *sound* as an adjective prevents the potential error in sentiment determination in such cases.

The contribution of the parse-tree information to the system performance was evaluated through the comparison between a system that uses valence shifters with uniform scope set to "all" (that is, spanning from the valence shifter forward or backward to the next punctuation mark) and more fine-grained scope definitions for different shifters (e.g., NP, ADJ, VP). Surprisingly, the contribution of fine-grained syntactic component was relatively small (1.5-2.5%) compared to the more coarse estimate of the valence shifter scope and this complex additional processing brought only small improvement to the system accuracy (Table 21[23]). The improvement was most marked on news, where accuracy increased by 2 percent points and precision — by 7 percent points. For other genres, the improvement was mostly related to higher recall, while precision remained the same or even declined slightly. This observation confirms that different genres require different approaches: movie and product reviews can be classified by simple weighted LBA, while news require a more

[23] The addition of parse information brought statistically significant improvement compared to the baseline for News and PRs at $\alpha = 0.01$, and for Blogs and Movie reviews at $\alpha = 0.025$. At the same time, the change was not statistically significant compared to the results with more coarse approximation of valence shifter scope for all genres, except Blogs ($\alpha = 0.025$).

	Ternary classification	Binary classification			
	Accuracy PNO	Accuracy PN	P	R	F
News					
weighted	51.1	61.9	71.9	86.1	78.4
weighted, valence shifters, no parse	54.3	62.8	76.4	82.2	79.2
weighted, valence shifters, with parse	54.4	64.7	83.9	77.1	80.4
Blogs					
weighted	58.4	63.5	78.6	80.8	79.7
weighted, valence shifters, no parse	56.9	59.5	75.8	78.5	77.1
weighted, valence shifters, with parse	58.4	62.0	76.8	80.8	78.7
Movie reviews					
weighted	n/a	57.3	64.9	88.4	74.8
weighted, valence shifters, no parse	n/a	57.5	67.2	85.5	75.3
weighted, valence shifters, with parse	n/a	58.1	66.8	87.0	75.6
Product reviews					
weighted	n/a	57.4	72.8	78.8	75.7
weighted, valence shifters, no parse	n/a	56.1	75.0	74.8	74.9
weighted, valence shifters, with parse	n/a	59.0	76.8	76.8	76.8

Table 21: Lexicon-based approach with valence shifters, role of parse information.

sophisticated system.

4.2.4 Evaluation

The validation of this approach requires the comparative evaluation on several domains and at different linguistic levels. The research on sentiment annotation is usually conducted at text [11, 94, 95, 104, 130, 132] or at sentence levels [46, 56, 63, 104, 156]. It should be noted that each of these levels presents a unique set of challenges in sentiment annotation. For example, it has been observed that texts often contain multiple opinions on different topics [130, 141], which makes assignment of the overall sentiment to the whole document problematic. On the other hand, each individual sentence contains a limited number of sentiment clues, which often negatively affects the accuracy and recall of annotation if that single sentiment clue encountered in the sentence was not learned by the system.

In this section, the results on sentence-level sentiment classification are compared to text-level sentiment classification for movie reviews and the classification of headlines for news[24] (Table 22) and to the results from the extant literature, in order to put them into perspective.

[24] The comparison is limited to news and movie reviews because there are no comparable datasets for other two genres — blogs and product reviews.

[25] The calculation of the accuracy of SVMs with movie review texts was not possible with the available resources given the large size of the corpus and corresponding feature space.

Genre	News		Movie reviews	
Unit	Sentence	Headline	Text	Sentence
Avg. length	25.6	6.5	690	21
Num of instances	800	1050	1522	1066
Naïve Bayes[25]	59.5	31.2	81.1	60.2
Baseline Lexicon-based	57.1	56.9	61.9	54.8

Table 22: Instance length impact on the performance of classifier performance (binary classification)

Table 22 shows that larger linguistic units (i.e., texts) are easier to classify. This is primarily due to a greater number of sentiment clues that these larger units contain. This contributed to system performance in two ways: First, greater number of sentiment clues makes decisions about the overall unit sentiment more reliable. This property is particularly important for the performance of lexicon-based approaches (LBAs). Second, for statistical classifiers, even a small number of training instances of large size allows the system to learn considerably more features and thus makes the probabilistic classifier more accurate. The impact of instance length is considerably greater for CBA where the gap is of 20% than for LBA, which maintains a relatively stable performance across different units.

The results of SemEval-2007 Affective Text task provide interesting insights into the comparative advantages and limitations of supervised machine learning and lexicon-based approaches. The participating system were evaluated by their precision, recall, and accuracy. Two measures of accuracy were used: Pearson correlation between fine-grained sentiment scores and scores assigned by the participating systems, and coarse-grained accuracy for binary positive vs. negative classification. We submitted two systems to this competition: one (ClaC) was using syntax-aware LBA, the other — ClaC-NB — was based on the Naïve Bayes classifier. ClaC-NB had the highest recall among the participating systems, while LBA ClaC demonstrated the highest precision and accuracy (both fine- and coarse- grained). Due to the small size of the development corpus (250 headlines), the participants had to use external resources in order to learn the features necessary for sentiment classification. In this setting, machine learning methods (a Naïve-Bayes-based CLaC-NB [5] (Table 22) and word-space model-based SICS [110] systems) had the highest recall at the cost of low precision and accuracy. This can be attributed to the lack of in-domain training data. On the other hand, lexicon-based unsupervised approaches (CLaC [5] and UPAR7 [25]) had the highest accuracy and precision, but their recall was low, probably due to the small size of headlines and a low number of sentiment clues per each headline.

4.3 Conclusions

Chapter 4 explored different aspects of the Lexicon-based approach to sentiment tagging — from lexicon acquisition to the study of factors that can improve the performance of a lexicon-based classifier. Knowledge-rich lexicon-based approaches to sentiment tagging received relatively little attention in the literature compared to the corpus-based methods. While the methods of acquisition of the lexicons and the performance of these lexicons in the text-level sentiment tagging have received some attention in the literature [56, 122, 38, 61, 63], substantial gaps in this domain still remained. This chapter thus contributes to the research on lexicon-based approaches to sentiment tagging by exploring the contribution of other kinds of information that can be added to lexicon-based systems: weights or scores and valence shifters. This chapter has explored the role of valence shifters and presented a comparative study of the impact of *weights*, *valence shifters*, *parse tree information* on performance of the lexicon-based approach. This part of the research builds upon the work on the sentiment tagging at the word-level, presented at the beginning of this chapter.

The research presented here contributes to the development of lexicon-based approach to sentiment tagging at two levels — word and sentence levels. Advancing the research at *word level*, where the identification of sentiment-bearing words represents the essential first step in the development of a lexicon-based system, Section 4.1 presented a novel approach to sentiment tagging of words and senses based on dictionary information. This approach contributes to the development of NLP and semantic tagging systems

- By generating a ***list of sentiment-bearing words*** from the entire WordNet dictionary with accuracy of the positive vs. negative classification at the word-level that is superior to approaches described in the extant literature.

- By developing ***annotation at the sense-level***, rather than word level, and introducing partial disambiguation of adjectives in texts based on the semantic category of the noun they modify, which opens a novel possibility for development of more accurate sentiment tagging systems.

- By exploring the ***structure of the semantic category of sentiment*** as a fuzzy set: the distance from the core of the category, as measured by Net Overlap Scores derived from multiple STEP runs, is shown to affect both the level of inter-annotator agreement and the system performance vs. human-annotated gold standard.

- By highlighting important ***considerations in system evaluation*** at the word level.

- First, this research emphasizes the importance of multiple runs for accurate evaluation of system performance, since system results vary dramatically depending on the composition of the seed list.
- Second, this study demonstrates that the centrality of a given word to the semantic category of sentiment, as measured by Net Overlap Score values, has an important effect on the accuracy of both automatic and human-made determination of the word's sentiment.
- Finally, the chapter calls for a reconsideration of the role of inter-annotator disagreement in the semantic annotation of words: the inter-annotator agreement was shown to fall as we proceed from the core of a fuzzy semantic category to its periphery and, hence, reflects, the fuzziness of the semantic category rather than a quality problem in human annotation.

The resulting list of words tagged with sentiment was used in the second part of the chapter as one of the inputs to the lexicon-based sentiment tagging at *sentence level.* The experiments conducted at this level sought to assess the quality of the obtained wordlist and the impact of different additional kinds of knowledge that a lexicon-based system can use: weights, valence shifters, and syntactic information. The findings from these experiments were that:

- The *scoring* of words with NOS, introduced in Section 4.1, improves the performance of the system.
- The addition of *valence shifter handling* marginally improves the performance of the lexicon-based approach.
- The experiments with more fine-grained scope approximation for valence shifters based on the dependency information did not produce the expected improvement in performance mostly due to the lack of reliable scope determination.
- A lexicon-based approach performs more consistently across different *genres and domains* than a corpus-based approach.

Chapter 5

Combined Approach: Corpus-based and Lexicon-based Methods in an Ensemble of Classifiers

Domain portability is one of the most important problems in the state-of-the-art sentiment tagging research. This chapter addresses this problem by proposing an ensemble of classifiers approach that combines the strengths of lexicon-based and corpus-based approaches using a precision-based vote weighting technique developed in this study. The first section of this chapter provides an overview of issues in system portability and a review of the literature on domain adaptation in sentiment analysis. The second part establishes a baseline for system evaluation by drawing comparisons of system performance across four different domains and genres - movie reviews, news, blogs, and product reviews. The final, third part of the chapter presents the system, composed of an ensemble of two classifiers — one trained on WordNet glosses and synsets and the other trained on a small in-domain training set.

5.1 The Problem of System Portability across Different Domains and Genres

The previous chapters have addressed corpus-based and sentiment based approaches to sentiment analysis at different levels (Chapters 3 and 4). The analysis presented there suggested that sentence and text-level sentiment classifiers that use standard machine learning techniques to learn and select features from labeled corpora work well in situations where large labeled corpora are available for training and validation, but they do not perform well when training data is scarce or when it comes from a different domain [11, 102], topic [102], or time period [102]. Given that these are very common real-world constraints on training data quality and availability, the problem of *system portability across different domains* becomes a serious issue for practical applications of corpus-based approaches to sentiment annotation.

5.1.1 System Portability Problem with a Corpus-based Classifier

Supervised statistical methods have been very successful in sentiment tagging of texts: on movie review texts they reach accuracies of 85-90% [11, 95]. These methods usually perform well when a large volume of labeled data from the same domain as the test set is available for training [11]. For this reason, most of the research on sentiment tagging using statistical classifiers was limited to product and movie reviews, where review authors usually indicate their sentiment in a form of a standardized score that accompanies the texts of their reviews.

The lack of sufficient data for training appears to be the main reason for the virtual absence of experiments with statistical classifiers in sentiment tagging at the sentence level. To our knowledge, the only work that describes the application of statistical classifiers (SVM) to sentence-level sentiment classification is [46][1]. The average performance of the system on ternary classification (positive, negative, and neutral) was between 0.50 and 0.52 for both average precision and recall. The results reported by Riloff et al. [104] for binary classification of sentences in a related domain of *subjectivity* tagging (i.e., the separation of sentiment-laden from neutral sentences) suggest that statistical classifiers can perform well on this task: the authors have reached 74.9% accuracy on the MPQA corpus [104].

Classifier selection

In order to illustrate the performance of different approaches in sentiment annotation at the text and sentence levels, I used a basic Naïve Bayes classifier. It has been shown that both Naïve Bayes and SVMs perform with similar accuracy on different sentiment tagging tasks [95]. These observations were confirmed with our own experiments with SVMs and Naïve Bayes (Table 4). I used the Weka package (http://www.cs.waikato.ac.nz/ml/weka/) with default settings.

[1] Recently, a similar task has been addressed by the Affective Text Task at SemEval-1 where even shorter units – headlines – were classified into positive, negative and neutral categories using a variety of techniques [121].

Classifier performance on different domains

A set of experiments presented here compares corpus-based classifier results on sentences using in-domain and out-of-domain training[2]. Table 23 shows that in-domain training, as expected, consistently yields superior accuracy than out-of-domain training across all four datasets: movie reviews (Movies), news, blogs, and product reviews (PRs). The numbers for in-domain trained runs are highlighted in bold. However, when the classifier trained on data from one domain is ported onto other domain, system results deteriorate substantially.

	Test Data			
Training Data	Movies	News	Blogs	PRs
Movies	**68.5**	55.2	53.2	60.7
News	55.0	**61.5**	56.3	57.4
Blogs	53.7	49.9	**63.9**	58.8
PRs	55.8	55.9	56.3	**76.9**

Table 23: Accuracy of SVM with unigram model without feature selection.

It is also interesting to note that, as shown in Chapter 3 (Table 4 *on sentences*), regardless of the domain used in system training and regardless of the domain used in system testing, unigrams tend to perform better than higher-order n-grams. This observation suggests that, given the constraints on the size of the available training sets, unigram-based systems may be better suited for sentence-level sentiment annotation.

5.1.2 Approaches to the System Portability Problem

As mentioned earlier, there are two alternatives to supervised machine learning that can be used to get around the system portability problem: on the one hand, *general lists* of sentiment clues/features can be acquired from domain-independent sources such as dictionaries or the Internet, on the other hand, unsupervised and weakly-supervised approaches can be used to take advantage of a *small number of annotated in-domain examples and/or of unlabelled in-domain data.*

The first approach, which uses general word lists automatically acquired from the Internet or from dictionaries, outperforms corpus-based classifiers when such classifiers use out-of-domain training data or when the training corpus is not sufficiently large to accumulate the necessary feature frequency information. But such general word lists were shown to perform worse than statistical models built on sufficiently large in-domain training sets of movie reviews [94].

[2] In this experiments SVM was used without feature selection.

The second approach seeks to address performance deficiencies of supervised machine learning methods with insufficient or out-of-domain training by using unsupervised or weakly-supervised feature learning. For instance, Aue and Gamon [11] proposed training on a small number of labeled examples and large quantities of unlabelled in-domain data. This system performed well even when compared to systems trained on a large set of in-domain examples: on feedback messages from a web survey on knowledge bases, Aue and Gamon [11] report 73.86% accuracy using unlabelled data compared to 77.34% for in-domain and 72.39% for the best out-of-domain training on a large training set.

Blitzer et al. [16] applied structural correspondence learning [17] to the task of domain adaptation for sentiment classification of product reviews. They showed that, depending on the domain, a small number (e.g., 50) of labeled examples allows adaptation of the model learned on another corpus to a new domain. However, they note that the success of such adaptation and the number of necessary in-domain examples depends on the similarity between the original domain and the new one. Similarly, Tan et al. [127] suggested to combine out-of-domain labeled examples with unlabelled ones from the target domain in order to solve the domain-transfer problem. They applied an out-of-domain-trained SVM classifier to label examples from the target domain and then retrained the classifier using these new examples. In order to maximize the utility of the examples from the target domain, these examples were selected using Similarity Ranking and Relative Similarity Ranking algorithms [127]. Depending on the similarity between domains, this method brought up to 15% gain compared to the baseline SVM.

Overall, the development of semi-supervised approaches to sentiment tagging is a promising direction of the research on system portability across different domains but so far, based on reported results, the performance of such methods is inferior to the supervised approaches with in-domain training and to the methods that use general word lists.

The sections that follow present a novel approach to the problem of system portability across different domains by developing a sentiment annotation system that integrates a corpus-based classifier with a lexicon-based system trained on WordNet. By adopting this approach, I sought to develop a system that relies on both general and domain-specific knowledge, as humans do when analyzing a text. The information contained in lexicographical sources, such as WordNet, reflects a lay person's general knowledge about the world, while domain-specific knowledge can be acquired through classifier training on a small set of in-domain data. The sections that follow present our system composed of an ensemble of two classifiers – one trained on WordNet glosses and synsets and the other trained on a small in-domain training set. Thus, in addition to the **supervised corpus-based approaches** presented earlier, where a classifier is trained on different amounts of labeled data (Chapter

3) and **unsupervised lexicon-based approach** that uses the list of words (unigrams) learned from WordNet or other lexicographic sources (Chapter 4), this study presents a **combined approach**, where corpus-based and lexicon-based systems are integrated into an ensemble of classifiers.

5.1.3 Ensemble of Classifiers Approach in the Literature

The strategy of integration of two or more systems in a single ensemble of classifiers has been actively used on different tasks within NLP. In sentiment tagging and related areas, Aue and Gamon [11] demonstrated that combining classifiers can be a valuable tool in domain adaptation for sentiment analysis. In the ensemble of classifiers, they used a combination of nine SVM-based classifiers deployed to learn unigrams, bigrams, and trigrams on three different domains, while the fourth domain was used as an evaluation set. Using an SVM meta-classifier trained on a small number of target domain examples to combine the nine base classifiers, they obtained a statistically significant improvement on out-of-domain texts from book reviews, knowledge-base feedback, and product support services survey data. No improvement occurred on movie reviews.

Pang and Lee [95] applied two different classifiers to perform sentiment annotation in two sequential steps: the first classifier separated subjective (sentiment-laden) texts from objective (neutral) ones and then the second classifier classified the subjective texts into positive and negative. das and Chen [31] used five classifiers to determine market sentiment on Yahoo! postings. Simple majority vote was applied to make decisions within the ensemble of classifiers and achieved accuracy of 62% on ternary in-domain classification.

Kennedy and Inkpen [61] proposed to combine machine-learning (SVM) with term counting approach and tested this ensemble of classifiers on movie reviews texts. The weights were assigned using two methods: using a simple weighted average of the scores given by the two base learners, and using a meta-classifier with meta-scores as features. The improvement over the baseline SVM with unigrams for 10-fold cross-validation experiments was 1.3%. Kennedy and Inkpen [61] suggested that the improvement may be due to the fact that the two classifiers did not make the same kind of errors. The results reported in the literature thus suggests that a combination of two classifiers may result in an improvement of classifier performance.

5.2 Integrating the Corpus-based and Dictionary-based Approaches

In this section I describe an approach that attempts to address the problem of system portability by integrating a corpus-based approach (CBA) (Chapter 3) with a lexicon-based approach (LBA) described in Chapter 4.

5.2.1 Theoretical Rationale

Since benefits from combining classifiers that always make similar decisions is minimal, the two (or more) base-learners should complement each other [2]. For this reason, a system based on a fairly different learning approach is more likely to produce a different decision under a given set of circumstances. Thus, the diversity of approaches integrated in the ensemble of classifiers was expected to have a beneficial effect on the overall system performance.

These considerations suggested that lexicon-based systems can produce greatest synergies with a classifier trained on small-set in-domain data. A lexicon-based approach capitalizes on the fact that dictionaries, such as WordNet [42], contain a comprehensive and domain-independent set of sentiment clues that exist in general English. A system trained on such general data, therefore, should be less sensitive to domain changes. This robustness, however, is expected to come at some cost, since some domain-specific sentiment clues may not be covered in the dictionary. Our hypothesis was, therefore, that a lexicon-based system will perform worse than an in-domain trained classifier but possibly better than a classifier trained on out-of domain data.

Overall, by attempting the integration of corpus-based and lexicon-based approaches, I sought to develop a system that relies on both general and domain-specific knowledge, as humans do when analyzing a text. The selection of these two classifiers for this system, thus, was theory-based.

5.2.2 Establishing the Baselines for the Lexicon-based and Corpus-based Learners

The baseline performance of the Lexicon-Based System (LBS) described in Chapter 4 is presented in Table 24, along with the performance results of the in-domain- and out-of-domain-trained SVM classifier.

Table 24 confirms the predicted pattern: the LBS performs with lower accuracy than *in-domain*-trained corpus-based classifiers, and with similar or better accuracy than the

	Movies	News	Blogs	PRs
Lexicon-based	58.1	64.7	62.0	59.0
SVM in-domain	68.5	61.5	63.9	76.9
SVM out-of-domain	55.8	55.9	56.3	60.7

Table 24: System accuracy on best runs on sentences.

corpus-based classifiers trained on *out-of-domain* data. Thus, the lexicon-based approach is characterized by a bounded but stable performance when the system is ported across domains. These performance characteristics of corpus-based and lexicon-based approaches prompt further investigation into the possibility to combine the portability of dictionary-trained systems with the accuracy of in-domain trained systems.

The section that follows describes the classifier integration and presents the performance results of the system consisting of an ensemble corpus-based and lexicon-based classifier and a precision-based vote weighting procedure.

5.2.3 The Classifier Integration Procedure and System Evaluation

The comparative analysis of the corpus-based and lexicon-based systems described in Chapters 3 and 4 revealed that the errors produced by the corpus-based system (CBS) and lexicon-based system (LBS) were to a great extent complementary (i.e., where one classifier makes an error, the other tends to give the correct answer). This provided further justification to the integration of corpus-based and lexicon-based approaches in a single system.

Table 25 below illustrates the complementarity of the CBS and LBS classifiers on the positive and negative categories. In this experiment, the corpus-based classifier was trained on 400 annotated product review sentences[3]. The two systems were then evaluated on a test set of another 400 product review sentences. The results reported in Table 25 are statistically significant at $\alpha = 0.01$.

	CBA	LBA
Precision positives	**89.3%**	69.3%
Precision negatives	55.5%	**81.5%**
Pos/Neg Precision	58.0%	72.1%

Table 25: Base-learners' precision on product reviews on test data.

Table 25 shows that the corpus-based system has a very good precision on those sentences that it classifies as positive but makes many errors on those sentences that it deems negative.

[3] The small training set explains relatively low overall performance of the CBS system.

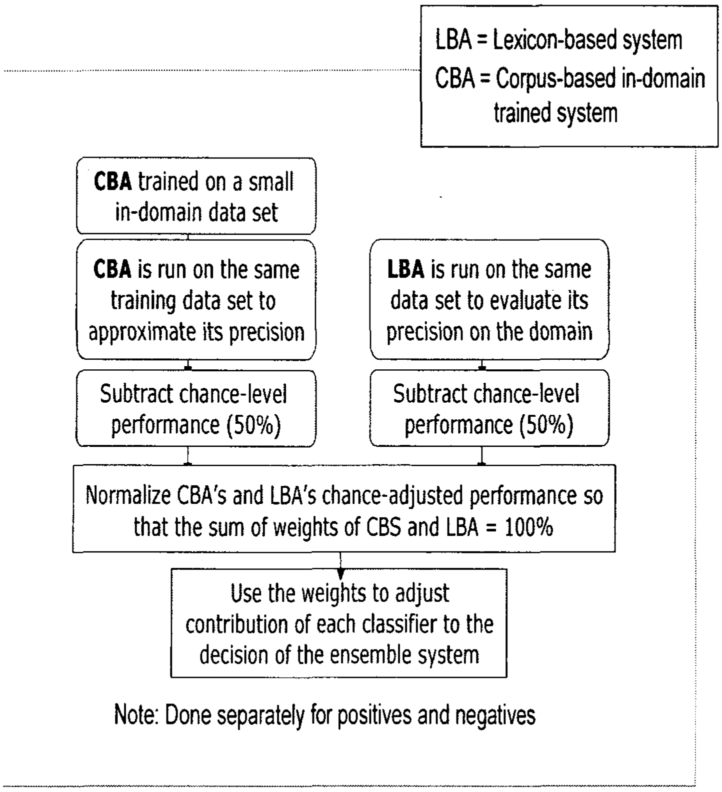
LBA = Lexicon-based system
CBA = Corpus-based in-domain trained system
CBA trained on a small in-domain data set
CBA is run on the same training data set to approximate its precision
LBA is run on the same data set to evaluate its precision on the domain
Subtract chance-level performance (50%)
Subtract chance-level performance (50%)
Normalize CBA's and LBA's chance-adjusted performance so that the sum of weights of CBS and LBA = 100%
Use the weights to adjust contribution of each classifier to the decision of the ensemble system
Note: Done separately for positives and negatives

Figure 10: Precision-based Voting Algorithm.

At the same time, the lexicon-based system has low precision on positives and high precision on negatives[4]. Such complementary distribution of errors produced by the two systems was observed on different data sets from different domains, which suggests that the observed distribution pattern reflects the properties of each of the classifiers, rather than the specifics of the domain/genre.

In order to take advantage of the observed complementarity of the two systems, the following procedure was used (Figure 10). First, a small set of in-domain data was used to train the CBS system. Then both CBS and LBS systems were run separately on the same training set, and for each classifier, the precision measures were calculated separately for those sentences that the classifier considered positive and those it considered negative. The chance-level performance (50%) was then subtracted from the precision figures to ensure that the final weights reflect by how much the classifier's precision exceeds the chance level. The resulting chance-adjusted precision numbers of the two classifiers were then normalized,

[4] These results are consistent with an observation in [61], where a lexicon-based system performed with a better precision on negative than on positive texts.

so that the weights of CBS and LBS classifiers sum up to 100% on positive and to 100% on negative sentences. These weights were then used to adjust the contribution of each classifier to the decision of the ensemble system. The choice of the weight applied to the classifier decision, thus, varied depending on whether the classifier scored a given sentence as positive or as negative. For example, if on the development set the CBS precision on negative sentences is 89.5% and LBS precision is 69.3%, the corresponding chance adjusted values will be 39.5 and 19.3, respectively. Then the weight for CBS will be 39.5/(39.5+19.5)=0.67 and for LBS 19.5/(39.5+19.5)=0.33. That means that when in the test set a sentence is labeled negative by CBS, the score assigned by this classifier is multiplied by 0.67, and when LBS considers a sentence negative its decision has a smaller weight, because its score is multiplied by 0.33.

The resulting system was then tested on a separate test set of sentences[5]. The small-set training and evaluation experiments with the system were performed on different domains using 3-fold validation.

The experiments conducted with the Ensemble system were designed to explore system performance under conditions of limited availability of annotated data for classifier training. For this reason, the numbers reported for the corpus-based classifier do not reflect the full potential of machine learning approaches when sufficient in-domain training data is available. Table 26 presents the results of these experiments by domain/genre. The results are statistically significant at $\alpha = 0.01$, except the runs on movie reviews where the difference between the LBS and Ensemble classifiers was significant at $\alpha = 0.05$.

		LBA	CBA	Ensemble
News	Acc	67.8	53.2	**73.3**
	F	0.8	0.7	**0.9**
Movies	Acc	54.5	53.5	**62.1**
	F	0.7	0.7	**0.8**
Blogs	Acc	61.2	51.1	**70.9**
	F	0.8	0.7	**0.8**
PRs	Acc	59.5	58.9	**78.0**
	F	0.8	0.7	**0.9**
Average	Acc	60.7	54.2	71.1
	F	0.8	0.7	0.8

Table 26: Performance of the ensemble classifier

Table 26 shows that the combination of two classifiers into an ensemble using the weighting technique described above leads to consistent improvement in system performance across

[5] The size of the test set varied in different experiments due to the availability of annotated data for a particular domain.

all domains/genres. In the ensemble system, the average gain in accuracy across the four domains was 16.9% relative to CBS and 10.3% relative to LBS. Moreover, the gain in accuracy and precision was not offset by decreases in recall: the net gain in recall was 7.4% relative to CBS and 13.5% vs. LBS. The ensemble system on average reached 99.1% recall. The F-measure has increased from 0.77 and 0.72 for LBS and CBS classifiers respectively to 0.83 for the whole ensemble system. The choice of the most complementary base learners and precision-based voting are the main advantages of the approach described here. It leads to a gain in accuracy ranging from 5 to 17 percent points over the base learners, compared to 1.3% reported by Kennedy and Inkpen [61] for a similar combination of machine learning and term-counting with a larger training set.

5.3 Discussion

The development of portable (i.e., domain- and genre-independent) sentiment determination systems poses a substantial challenge for researchers in NLP and artificial intelligence. The results presented in this study suggest that the integration of two fairly different classifier learning approaches in a single ensemble of classifiers can yield substantial gains in system performance on all measures. The most substantial gains occurred in recall, accuracy, and F-measure.

This study permits to highlight a set of factors that enable substantial performance gains with the ensemble of classifiers approach. Such gains are most likely when (1) the errors made by the classifiers are complementary, i.e., where one classifier makes an error, the other tends to give the correct answer, (2) the classifier errors are not fully random and occur more often in a certain segment (or category) of classifier results, and (3) there is a way for a system to identify that low-precision segment and reduce the weights of that classifier's results on that segment accordingly. The two classifiers used in this study – corpus-based and lexicon-based – provided an interesting illustration of potential performance gains associated with these three conditions. The use of precision of classifier results on the positives and negatives proved to be an effective technique for classifier vote weighting within the ensemble.

5.4 Conclusions

This study contributes to the research on sentiment tagging, domain adaptation, and the development of ensembles of classifiers (1) by proposing a novel approach for sentiment determination at sentence level and delineating the conditions under which greatest synergies among combined classifiers can be achieved, (2) by describing a precision-based technique for assigning differential weights to classifier results on different categories identified by the classifier (i.e., categories of positive vs. negative sentences), and (3) by proposing a new method for sentiment annotation in situations where the annotated in-domain data is scarce and insufficient to ensure adequate performance of the corpus-based classifier, which still remains the preferred choice when large volumes of annotated data are available for system training.

Among the most promising directions for future research in the direction laid out in this dissertation is the deployment of more advanced classifiers and feature selection techniques that can further enhance the performance of the ensemble of classifiers. The precision-based vote weighting technique may prove to be effective also in situations, where more than two classifiers are integrated into a single system. I expect that these more advanced ensemble-of-classifiers systems would inherit the benefits of multiple complementary approaches to sentiment annotation and will be able to achieve better and more stable accuracy on in-domain, as well as on out-of-domain data.

Chapter 6

Discussion and Conclusion

This dissertation addressed the problem of sentence-level sentiment tagging and the issue of domain portability, which poses a substantial problem in different areas of NLP. This thesis attempted a systematic study of two major approaches to sentiment tagging — corpus-based and lexicon-based — and proposed a combined approach that attempts the integration of the benefits of these two methods.

Section 6.1 below provides a brief overview of approaches developed and evaluated in this study and highlights the main findings and the thesis contribution to the research on automatic sentiment annotation and semantic tagging, while Section 6.2 outlines most prominent directions for future research that stem from this thesis.

6.1 Main Findings and Contributions of the Thesis

Since the research presented in this thesis has been conducted at different levels of analysis (words, sentences, and texts), the sections below proceed from the discussion of the word-level sentiment tagging method to discussion of the sentence- and text-level approaches, highlighting most important findings and contributions.

6.1.1 Word-level Sentiment Tagging

The development of a method for acquisition of sentiment-laden words was motivated by a need to develop a list of words that can then be used as feature set by a classifier in analysis of sentences and texts (lexicon-based approach).

Until recently, word-level sentiment was assigned based on corpus information [52, 130, 132] or on WordNet relations [60, 62, 56]. This thesis described a new approach to word and sense level sentiment tagging that uses not only semantic relations but also the important and often neglected information contained in dictionary glosses. This approach is linguistically motivated and has an accuracy that is better or comparable to the performance of other approaches to word-level sentiment tagging [38, 39, 60]. The main advantage of the approach developed in this dissertation is the concept of Net Overlap Score that is interpreted as a measure of the degree of word's membership in the fuzzy category of sentiment. The scores are obtained based on the results of multiple runs of the algorithm with small non-intersecting seed lists. Starting from a seed word, the algorithm assigns the sentiment to other words that are related to this seed word through semantic relations such as synonymy and antonymy, and through the presence of the seed word in the glosses of other synsets in WordNet. Some words appear in the results of only one such run which indicates that they are related to only one other sentiment-bearing word, other words are brought by many runs, which indicates that they are related to many other sentiment-laden words. Word's Net Overlap Score (NOS) that is computed based on the number of runs that assigned positive or negative sentiment to this word, reflect the number of these ties and thus the degree of centrality of a given word to the fuzzy category of sentiment.

The thesis empirically demonstrated that NOS scores are directly related to the level of inter-annotator agreement on word labels and to the accuracy of automatic sentiment tagging: higher NOS values correlate with higher annotator agreement and higher system accuracy. NOS values also proved to be a useful addition to a lexicon-based system for sentence-level sentiment tagging: they improved system accuracy on all genres. At the present time, NOS score developed here is the only automatically generated sentiment scoring feature that can be used as a measure of word"s centrality to the category of sentiment and that provides notable improvements in the performance of sentence-level sentiment taggers.

In empirical testing of the Sentiment Tag Extraction System (STEP) developed in this study for sentiment-laden lexica acquisition, the system's 88% accuracy in *binary* classification was superior to the accuracy reported in the literature for other systems run on large corpora [131, 52]. Turney and Littman [131] report 76.06% accuracy for experimental runs on $3,596$ sentiment-marked GI words from different parts of speech using a 2x10^9 corpus to compute point-wise mutual information between the GI words and 14 manually selected positive and negative paradigm words. When they used the entire World Wide Web, their accuracy was 82.84%. The accuracy of the approach described in this thesis on *ternary* classification of adjectives is better than that reported by Kim and Hovy [66]: average STEP accuracy in the 58 runs was 71.2% (Table 12), while Kim and Hovy [66] report accuracy of 69.1%. These numbers are, however, not directly comparable due to the differences in the gold standard.

The results obtained for STEP were also significantly better than those obtained by Esuli and Sebastiani [39] on binary classification and similar to [39] in ternary classification. In their experiments with binary classification, the best results on GI's positive and negative words (neutrals excluded) was 83%, on ternary classification they reached the accuracy of 66%.

6.1.2 Sentence-level Sentiment Tagging

Most research on sentiment concentrates on text-level analysis, where labeled data is relatively easy to obtain (e.g., movie reviews with user-made ratings). It has been demonstrated [130, 141, 95], however, that for a variety of genres, texts combine segments with different sentiment: product reviews can contain contrasting opinions on different aspects/features of the product, movie reviews are a mixture of objective and subjective sentences, and news texts present balanced views on the issues. Subjectivity analysis has worked with phrases and sentences for several years already [149, 106, 146, 104], but in sentiment tagging, sentence-level research is practically non-existent. Only a few studies in this domain have been performed to date [11, 56]. The study presented in this dissertation addresses this gap by conducting a comprehensive study of different approaches to sentiment tagging at sentence level and exploring the factors that can influence system performance.

First, based on literature on text-level sentiment analysis, we have identified two major approaches to sentiment tagging: corpus-based (CBA) and lexicon-based (LBA), and a list of factors that could potentially influence the performance of each. In order to ensure reliability and generalizability of the findings, they were tested on four different genres: news, blogs, movie reviews and product reviews. This study confirmed that CBA performance depends on the training corpus size, n-gram size, classifier algorithm, domain/genre, and feature selection method. It was shown that sentence-level sentiment annotation is more difficult than text-level sentiment tagging for the same domain, and in general smaller language units (sentences) are harder to classify than larger ones (texts). The experiments also demonstrated that, similar to text-level sentiment annotation [11], sentence-level corpus-based sentiment tagging is highly domain dependent: its performance differs depending on the domain/genre and deteriorates significantly when training and testing domains are different.

For lexicon-based approach (LBA), other factors proved to be relevant: the weights assigned to words in the lexicon, the use of valence shifters and of syntactic information. NOS-based weights were shown to contribute to system accuracy. For the valence shifters, the findings reported in the literature on text-level sentiment are inconsistent: Kennedy and Inkpen [61] report a small improvement in classification accuracy after the addition on valence shifters, while Dave et al. [32] observed a decrease in performance of their classifier when they added negated adjectives (like *NOT_good*) to the list of features. The experiments with a lexicon-based system (LBA) conducted for this dissertation showed that, for sentence-level sentiment annotation, the effect of addition of valence shifter handling depends on the genre: valence shifters and related syntactic information were useful for news sentences, had almost no effect on classifier performance on movie and product reviews, and had a small negative impact on blogs. These differences can be attributed to the differences in genres: news abound in valence shifters and the ability to handle these elements increases system's accuracy and precision. The reviews of both types have much fewer valence shifters and, therefore, are practically indifferent to this system component. Blogs, which often contain agrammatical phrases, do not lend themselves well to complex syntactic processing necessary for valence shifter handling. For this reason, the performance of LBA on blogs slightly deteriorates with addition of valence shifter handling.

Although lexicon-based approach had lower performance than corpus-based method with in-domain training, the comparative experiments conducted in this study revealed two important advantages of LBA over CBA: LBA system's performance is much more consistent across domains and it does not require training data. These two properties make it much more portable and domain-independent than CBA. The development and testing of approaches that are complementary to the "mainstream" corpus based methods represents an important contribution of this thesis and a promising direction for future research. The development of such approaches can provide solutions to the problems and shortcomings encountered with more traditional methods and thus can further advance the research in NLP.

6.1.3 The Combined Approach in Sentiment Tagging

The observed complementarity of CBA and LBA prompted the development of a novel approach to domain adaptation in sentiment tagging that would combine the benefits of the two approaches.

The problem of domain portability is well known in sentiment tagging and other text classification tasks. It has attracted a considerable attention in sentiment analysis in the recent years as the performance of out-of-domain trained statistical classifiers is fairly low [11, 102, 16]. Most solutions proposed to date either use bootstrapping [11] or semi-supervised learning that combines training on both labeled and unlabelled data [49] or on in- and out-of-domain training examples [16]. All the approaches described in the literature seek a solution within CBA paradigm. To the contrary, the method described in this dissertation relies on the complementarity and relative merits of LBA and CBA in order to improve classifier performance when little training data is available. The comparison of CBA and LBA demonstrated that these two classifiers complement each other: LBA is domain-independent, general method, while CBA has high in-domain accuracy. It is also important to note that the classifiers combined into an ensemble do not have to be the best performing ones, on the contrary, the complementarity of errors should be the primary consideration when choosing specific base-learners to be put together.

Moreover, I have observed that the baseline corpus-based approach provides higher precision on positive sentences, while baseline lexicon-based approach performed better on negative sentences. Based on this observation, a new method of combining these two base-learners in an ensemble of two classifiers was developed in this dissertation. It dynamically learns the weights for each classifier contribution using the same small set of labeled data that was used to train the corpus-based component. This approach brought an increase in accuracy between 10% and 16.9% relative to the performance of LBA and CBA base-learners taken individually. Its performance is inferior only to an in-domain trained SVM with feature-selection, but compared to that method, the ensemble classifier requires four to five times less training data to achieve a comparable performance level.

6.2 Directions for Future Research

Several directions for future NLP research stem from this study. First, one the most promising directions for future research in the direction laid out in this thesis is the optimization of the sentiment tagging systems through the development of more advanced classifiers and feature selection techniques that can further enhance the performance of the ensemble of classifiers. The precision-based vote weighting technique may prove to be effective also in situations, where more than two classifiers are integrated into a single system. We expect that these more advanced ensemble-of-classifiers systems will inherit the benefits of multiple complementary approaches to sentiment annotation and will be able to achieve better and more stable accuracy on in-domain, as well as on out-of-domain data.

Second, the research presented here demonstrated a high degree of complementarity between corpus-based and lexicon-based system, which can be exploited in other domains, outside the category of sentiment and sentiment tagging. The precision-based voting technique described here enables the ensemble of classifiers to benefit from the strengths of both corpus-based and lexicon-based systems.

The experiments with in- and out-of-domain training for supervised machine-learning approach to sentiment classification indicate that, depending on the domain/genre similarity, the effects of the out-of-domain training may be more or less strong. Further research into the impact of corpus similarity on the performance of corpus-based approaches may provide useful insights into the ways to select most useful out-of-domain datasets when in-domain corpora does not exist or is too small.

By taking into account these observations, future research can enhance accuracy and rigor of system evaluations and system performance comparisons.

In addition to these major theory-driven directions for further research, several more specific opportunities for further research and system development in the domain of automatic sentiment annotation stem from this study:

- Holder and target/topic detection for more precise sentiment annotation,
- Combination of sentiment annotation systems with other applications, such as summarization, Question Answering, Information Retrieval, etc. and
- Addition of textual connectors and discourse modifiers to the features used by lexicon-based approach in addition to valence shifter handling.

Overall, automatic sentiment annotation at different linguistic levels (words, sentences, texts) represents not only an area of important practical relevance (e.g., MPQA, summarization, public opinion studies, etc.), but also an important domain where new approaches to semantic tagging and text analysis can be tried, tested, and refined for applications in other areas, outside of sentiment research. Further research thus should address automatic annotation of other semantic categories and a broader scope of practical applications.

Bibliography

[1] Khurshid Ahmad, Lee Gillam, and David Cheng. Sentiments on the Grid: Analysis of Streaming News and Views. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, pages 2517–2520, Genova, Italy, May 2006.

[2] Ethem Alpaydin. *Introduction to Machine Learning.* The MIT Press, Cambridge, MA, 2004.

[3] Alina Andreevskaia and Sabine Bergler. Mining WordNet For a Fuzzy Sentiment: Sentiment Tag Extraction From WordNet Glosses. In *Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006.

[4] Alina Andreevskaia and Sabine Bergler. Sentiment Tagging of Adjectives at the Meaning Level. In L. Lamontagne and M. Marchard, editors, *The 19th Canadian Conference on Artifical Intelligence 2006*, volume 4013 of *LNAI*, pages 336–346. Springer, 2006.

[5] Alina Andreevskaia and Sabine Bergler. CLaC and CLaC-NB: Knowledge-based and Corpus-based Approaches to Sentiment Tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 117–120, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] Alina Andreevskaia and Sabine Bergler. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of Association for Computational Linguistics / Human Language Technology (ACL-08: HLT)*, pages 290–298, Columbus, Ohio, 2008.

[7] Alina Andreevskaia, Sabine Bergler, and Monica Urseanu. All Blogs are not made Equal. In *International Conference on Weblogs and Social Media (ICWSM-2007)*, Boulder, Colorado, March 2007.

[8] Blake Andrew, Lori Young, and Stuart Soroka. Back to the Future: Press Coverage of the 2008 Canadian Election Campaign Strikes both Familiar and Unfamiliar Notes. *Policy Opinions*, 29(10):78–84, 2008.

[9] Shlomo Argamon, Kenneth Bloom, Andrea Esuli, and Fabricio Sebastiani. Automatically Determining Attitude Type and Force for Sentiment Analysis. In *Proceedings of the 3rd Language and Technology Conference (LTC'07)*, pages 369–373, Poznan, Poland, October 2007.

[10] Laurent Audibert. Word sense disambiguation criteria: a systematic study. In *Proceedings of 20th International Conference on Computational Linguistic (COLING-04)*, pages 910–916, Geneva, Switzerland, August 2004.

[11] Anthony Aue and Michael Gamon. Customizing Sentiment Classifiers to New Domains: a Case Study. In *RANLP-05, the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, September 2005.

[12] Anne Banfield. *Unspeakable Sentences: Narration and Representation in the Language of Fiction.* Routledge and Kegan Paul, Boston, Massachusets, USA, 1982.

[13] Marco Baroni and Stefano Vegnaduzzo. Identifying Subjective Adjectives through Web-based Mutual Information. In Ernst Buchberger, editor, *Proceedings of Konferentz zur Verarbeitung Naturlicher Sprache (KONVENS)*, pages 17–24, Vienna, Austria, September 2004.

[14] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic Extraction of Opinion Propositions and their Holders. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 125–139. Springer, 2006.

[15] Larry Blair, Assad Jaharria, Stepehn Lewis, Tomohiro Oda, Christoph Reichenbach, Jeff Rueppel, and Franco Salvetti. Impact of Lexical Filtering on Semantic Orientation. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.* Springer, 2006.

[16] John Blitzer, Mark Drezde, and Fernando Pereira. Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification. In *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 440 – 447, Prague, Czech Republic, 2007.

[17] John Blitzer, McDonald Rayan, and Fernando Pereira. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 5th Conference on Empirical Methods in Natural Language Processing (EMNLP-06)*, pages 120–128, Syndey, Australia, 2006.

[18] Wendy Boswell. Opinmind, a Blog Search Engine. `http://websearch.about.com/-od/enginesanddirectories/a/opinmind.htm` retreived January 6, 2009.

[19] M.M. Bradley and P.J. Lang. *Affective norms for English words (ANEW)*. The NIMH Center for the Study of Emotion and Attention,University of Florida, Gainesville, FL, 1999.

[20] Eric Breck, Yejin Choi, and Claire Cardie. Identifying Expressions of Opinion in Context. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-2007)*, Hyderebad, India, 2007.

[21] Eric Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[22] Rebecca Bruce and Janyce Wiebe. Recognizing Subjectivity: A Case Study of Manual Processing. *Natural Language Engineering*, 5(2):187–205, 2000.

[23] R.S. Burt. Models of network structure. *Annual Review of Sociology*, 6:79–141, 1980.

[24] Pimwadee Chaovalit and Lina Zhou. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. In *Proceedings of HICSS-05, the 38th Hawaii International Conference on System Sciences*, page 112c, Big Island, Hawaii, USA, January 2005.

[25] François-Régis Chaumartin. UPAR7: A Knowledge-based System for Headline Sentiment Tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 422–425, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[26] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Vancouver, Canada, 2005.

[27] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. *Lexical Aquisition: Exploiting On-Line Resources to Build a Lexicon*, chapter Using Statistics in Lexical Analysis, pages 115–164. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[28] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[29] comScore/the Kelsey group. Online Consumer-Generated Reviews have Significant Impact on Offline Purchase Behavior. Press Release, `http://www.comscore.com/press/release.asp?press=1928`, November 2007.

[30] Hang Cui, Vibhu Mittal, and Mayur Datar. Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-2006)*, pages 443–444, Boston, Massachusets, USA, July 2006.

[31] Sanjiv R. Das and Mike Y. Chen. Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web. In *Asia Pacific Finance Association Annual Conference (APFA01)*, Bangkok, Thailand, July 2001.

[32] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of WWW '03, the 12th International Conference on World Wide Web*, pages 519–528, Budapest, Hungary, May 2003. ACM.

[33] Ann Devitt and Khurshid Ahmad. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984 – 991, Prague, Czech Republic, June 2007.

[34] Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19:61–74, 1993.

[35] Kathleen T. Durant and Michael D. Smith. *Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection.* Springer, Berlin / Heidelberg, 2007.

[36] P. Ekman. Facial expression of emotion. *American Psychologist*, 48:384–392, 1993.

[37] C. Elliot. *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System.* PhD thesis, Northwestern University, Institute for the Learning Sciences, 1992.

[38] Andrea Esuli and Fabrizio Sebastiani. Determining the Semantic Orientation of Terms Through Gloss Analysis. In *Proceedings of CIKM-05, the 14th ACM SIGIR Conference on Information and Knowledge Management*, pages 617–624, Bremen, Germany, November 2005.

[39] Andrea Esuli and Fabrizio Sebastiani. Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics*, pages 193–200, Trento, Italy, April 2006.

[40] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, pages 417–422, Genova, Italy, May 2006.

[41] Andrea Esuli and Fabrizio Sebastiani. PageRanking WordNet Synsets: An Application to Opinion Mining. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, pages 424 –431, Prague, Czech Republic, June 2007.

[42] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[43] Jeremy Fletcher and Jon Patrick. Evaluating the Utility of Appraisal Hierarchies as a Method for Sentiment Classification. In *Proceedings of Australian Language Technology Workshop 2005*, pages 134–142, Sydney, Australia, December 2005.

[44] George Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[45] Michael Gamon. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. In *Proceeding of COLING-04, the 20th International Conference on Computational Linguistics*, pages 841–847, Geneva, Switzerland, August 2004.

[46] Michael Gamon and Anthony Aue. Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 57–64, Ann Arbor, Michigan, USA, July 2005.

[47] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining Customer Opinions from Free Text. In *Proceedings of IDA-05, the 6th International Symposium on Intelligent Data Analysis*, volume 3646 of *Lecture Notes in Computer Science*, pages 121–132, Madrid, ES, 2005. Springer-Verlag.

[48] Michel Généreux, Thierry Poibeau, and Moshe Koppel. Sentiment Analysis using Automatically Labelled Financial News. In *Proceedings of LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology*, Marrakech, Morocco, May 2008.

[49] Andrew Goldberg and Jerry Zhu. Seeing Stars when there aren't Many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization. In *Proceedings of the HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*, pages 45–52, Rochester, new-York, USA, April 2006.

[50] Gregory Grefenstette, Yan Qu, David A. Evans, and James G. Shanahan. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 93–107. Springer Verlag, 2006.

[51] Vasileios Hatzivassiloglou and Kathleen B. McKeown. A Quantitative Evaluation of Linguistic Tests for the Automatic Prediction of Semantic Markedness. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 197–204, M.I.T, Cambridge, Massachusetts, June 1995. Association for Computational Linguistics.

[52] Vasileios Hatzivassiloglou and Kathleen B. McKeown. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, pages 174–181, Madrid, Spain, July 1997. Association for Computational Linguistics.

[53] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pages 299–305, Saarbrücken, Germany, July–August 2000. Morgan Kaufman.

[54] Marti Hearst. *Text-based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval)*, chapter Direction-based Text Interpretation as

an Information Access Refinement, pages 257–274. Lawrence Erlbaum Associates, Mahwah, NJ, 1992.

[55] J. A. Horrigan. Online Shopping. Pew Internet & American Life Project Report, 2008.

[56] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-04)*, pages 168–177, Seattle, Washington, USA, August 2004.

[57] Matthew Hurst and Kamal Nigam. Retrieving Topical Sentiments from Online Document Collection. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 265–280. Springer Verlag, 2006.

[58] Nancy Ide. Making Senses: Bootstrapping Sense-tagged Lists of Semantically-Related Words. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings*, volume 5449 of *LNCS*, pages 13–27. Springer, 2006.

[59] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[60] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. Using WordNet to measure semantic orientation of adjectives. In *4th International Conference on Language Resources and Evaluation (LREC 2004)*, volume IV, pages 115–118, Paris, 2004. European Language Resources Association.

[61] Alistair Kennedy and Diana Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, 2006.

[62] Soo-Min Kim and Eduard Hovy. Determining the Sentiment of Opinions. In *Proceedings COLING-04, the 20th International Conference on Computational Linguistics*, pages 1367–1373, Geneva, Switzerland, August 2004.

[63] Soo-Min Kim and Eduard Hovy. Automatic Detection of Opinion Bearing Words and Sentences. In *Companion Volume to the Proceedings of IJCNLP-05, the Second*

International Joint Conference on Natural Language Processing, pages 61–66, Jeju Island, Korea, October 2005.

[64] Soo-Min Kim and Eduard Hovy. Identifying Opinion Holders for Question Answering in Opinion Texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*, Pittsburgh, Pensylvannia, USA, 2005.

[65] Soo-Min Kim and Eduard Hovy. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 483–490, Sydney, Australia, July 2006. Association for Computational Linguistics.

[66] Soo-Min Kim and Eduard Hovy. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *Proceedings of the ACL/COLING Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics.

[67] Soo-Min Kim and Eduard Hovy. Identifying and Analyzing Judgment Opinions. In *Proceedings of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL-2006)*, pages 200–207, New-York, USA, June 2006.

[68] Soo-Min Kim and Eduard Hovy. Crystal: Analyzing Predictive Opinions on the Web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064, Prague, Czech Republic, June 2007.

[69] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1065–1074, Prague, Czech Republic, June 2007.

[70] Moshe Koppel and Jonathan Schler. The Importance of Neutral Examples for Learning Sentiment. *Computational Intelligence*, 22(2):100–116, May 2006.

[71] Claus Krippendorf. *Content Analysis: An Introduction to Its Methodology.* Sage, Thousand Oaks, CA, 2004.

[72] 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.

[73] Gilly Leshed and Joseph 'Jofish' Kaye. Understanding How Bloggers Feel: Recognizing Affect in Blog Posts. In *Proceedings of the 2006 ACM Conference on Human Factors in Computing Systems (CHI-2006)*, pages 1019–1024, Montreal, Quebec, Canada, April 2006.

[74] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the International Conference on Computational Linguistics / Conference of the Association for Computational Linguistics (COLING-ACL-98)*, pages 768–774, Montreal, Quebec, Canada, August 1998.

[75] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which Side are You on? Identifying Perspectives at the Document and Sentence Levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-06)*, pages 109–116, New York, US, June 2006. Association for Computational Linguistics.

[76] Hugo Liu, Henry Lieberman, and Ted Selker. A Model of Textual Affect Sensing using Real-world Knowledge. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI-03)*, pages 125–132, Miami, Florida, USA, January 2003.

[77] Bernardo Magnini and Gabriela Cavalia. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

[78] Christopher Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* The M.I.T Press, Cambridge, Massachusetts, USA, 1999.

[79] Yi Mao and Guy Lebanon. Sequential Models for Sentiment Prediction. In *Proceedings of the ICML Workshop: Learning in Structured Output Spaces Open Problems in Statistical Relational Learning Statistical Network Analysis: Models, Issues and New Directions*, Pittsburg, Pennsylvania, USA, June 2006.

[80] J.R. Martin and P.R.R. White. *The Language of Evaluation: Appraisal in English.* Palgrave, London, 2005.

[81] C. Mcdonald and I. Ounis. The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow, 2006.

[82] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 432–439, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[83] Rada Mihalcea. Word Sense Disambiguation using Pattern Learning and Automatic Feature Selection. *Natural Language Engineering*, 1(1):1–15, 2002.

[84] Rada Mihalcea and Hugo Liu. A Corpus-based Approach to Finding Happiness. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs, AAAI Technical report SS-06-03*, Stanford, California, USA, March 2006.

[85] Rada Mihalcea and Dan I. Moldovan. eXtended WordNet: Progress Report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100, Pittsburg, Pennsylvania, USA, June 2001.

[86] Gilad Mishne. Experiments with Mood Classification in Blog Posts. In *Proceedings of Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005*, 2005.

[87] Gilad Mishne and Natalie Glance. Predicting Movie Sales from Blogger Sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford, California, USA, 2006.

[88] Rahman Mukras. A Comparison of Machine Learning Techniques Applied To Sentiment Classification. Master's thesis, University of Sussex, Falmer, Brighton, 2004.

[89] Matthijs Mulder, Anton Nijholt, Marten den Uyl, and Peter Terpstra. A Lexical Grammatical Implementation of Affect. In *Proceedings of TSD-04, the 7th International Conference Text, Speech and Dialogue*, volume 3206 of *Lecture Notes in Computer Science*, pages 171–178, Brno, Czech Republic, September 2004. Springer.

[90] Tony Mullen and Nigel Collier. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 412–418, Barcelona, Spain, July 2004.

[91] J. Zvi Namewirth and Robert P. Weber. *Dynamics of Culture.* Allen and Unwin, Winchester MA, 1987.

[92] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, New York, 1988.

[93] Bo Pang and Lilian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retreival*, 2(1–2):1–135, 2008.

[94] Bo Pang, Lilian Lee, and Shrivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86. AAAI, 2002.

[95] Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 271–278, Barcelona, Spain, July 2004.

[96] Bo Pang and Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43nd Meeting of the Association for Computational Linguistics (ACL-05)*, pages 115–124, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

[97] R.W. Picard. *Affective Computing.* M.I.T. Press, Boston, Massachusets, USA, 1997.

[98] John Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Learning.* M.I.T. Press, Cambridge, Massachusets, USA, 1998.

[99] Livia Polanyi and Annie Zaenen. Contextual Valence Shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 1–10. Springer Verlag, 2006.

[100] Ana-Maria Popescu and Oren Etzioni. Extracting Product Features and Opinions from Reviews. In *Proceedings of the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP-05)*, pages 339–346, Vancouver, B.C., Canada, October 2005.

[101] John Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[102] Jonathon Read. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL-2005 Student Research Workshop*, pages 49–54, Ann Arbor, Michigan, USA, June 2005. Association for Computational Linguistics.

[103] Ellen Riloff and R. Johns. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 1044–1049, Orlando, Florida, USA, July 1999.

[104] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. Feature Subsumption for Opinion Analysis. In *Proceedings of the 5th Conference on Empirical Methods in Natural Language Processing (EMNLP-06)*, pages 440–448, Sydney, Australia, July 2006. Association for Computational Linguistics.

[105] Ellen Riloff and Janyce Wiebe. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 105–112, Sapporo, Japan, July 2003.

[106] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the 7th Conference on Natural Language Learning (CONLL-03)*, pages 25–32, Edmonton, CA, May-June 2003.

[107] Peter Mark Roget. *Roget's Thesaurus of English Words and Phrases.* T. Y. Crowell Co., 1911. Project Gutenberg$^{\mathrm{TM}}$etext.

[108] N. T. Roman, P. Piwek, and A. M. B. R Carvalho. Politeness and Bias in Dialogue Summarization: Two Exploratory Studies. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 171–183. Springer, 2006.

[109] W. Sack. On the Computation of Point of View. In *Proceedings of the National Conference of Artificial Intelligence (AAAI-94*, page 1448, July-August 1994.

[110] Magnus Sahlgren, Jussi Karlgren, and Gunnar Eriksson. SICS: Valence Annotation based on Seeds in Word Space. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 296–299, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[111] Franco Salvetti, Stephen Lewis, and Christoph Reichenbach. Impact of Lexical Filtering on Overall Polarity Identification. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 303–315. Springer Verlag, 2006.

[112] Gün R. Semin and Klaus Fiedler. The Cognitive Functions of Linguistic Categories in Describing Persons: Social Cognition and Language. *Journal of Personality and Social Psychology*, 54:558–568, 1988.

[113] Erick Shonfeld. Summize: A Sentiment Engine For The "Reviewosphere". *TechCrunch*, 2007. `http://www.techcrunch.com/2007/12/17/-summize-a-sentiment-engine-for-the-reviewosphere/`.

[114] Benjamin Snyder and Regina Barzilay. Multiple Aspect Ranking using the Good Grief Algorithm. In *Proceedings of North American chapter of the Association for Computational Linguistics annual meeting (NAACL-2007)*, pages 300–307, Rochester, New York, USA, April 2007.

[115] Ian Soboroff and Donna Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of The Twelfth Text Retrieval Conference*, page 38, Gaithersburg, Maryland, 2003.

[116] Marina Sokolova, Stan Szpakowicz, and Vivi Nastase. Using Language to Determine Success in Negotiations: A Preliminary Study. In *Proceedings of the 17th Canadian Conference on Artificial Intelligence (Canadian AI'2004)*, volume 3060 of *LNCS*, pages 449–453, London, Ontario, Canada, May 2004. Springer.

[117] Ellen Spertus. Smokey: Automatic Recognition of Hostile Messages. In *Proceedings of IAAI-97, the 9th Conference on Innovative Application of Artificial Intelligence*, pages 1058–1065, Providence, US, 1997.

[118] P.J. Stone, D.C. Dumphy, M.S. Smith, and D.M. Ogilvie. *The General Inquirer: a Computer Approach to Content Analysis.* M.I.T. studies in comparative politics. M.I.T. Press, Cambridge, MA, 1966.

[119] Veselin Stoyanov, Claire Cardie, Diane Litman, and Janyce Wiebe. Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 77–89. Springer, 2006.

[120] Carlo Strapparava and Rada Mihalcea. SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[121] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[122] Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: an Affective Extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1083–1086, Lisbon, Portugal, May 2004.

[123] Pero Subasic and Alison Huettner. Affect Analysis of Text Using Fuzzy Typing. *IEEE Transactions on Fuzzy Systems (IEEE-FS)*, 9:483–496, 2001.

[124] K. Sweeney and C. Whissell. A dictionary of Affect in Language: I. Establishment and Preliminary Validation. *Perceptual and Motor Skills*, 59(3):695–698, 1984.

[125] M. Taboada, C. Anthony, and K. Voll. Methods for Creating Semantic Orientation Databases. In *Proceeding of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 427–432, Genoa, Italy, May 2006.

[126] Maite Taboada and Jack Grieve. Analyzing Appraisal Automatically. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.*, pages 427–432. Springer Verlag, 2006.

[127] Songbo Tan, Gaowei Wu, Huifeng Tang, and Xueqi Cheng. A Novel Scheme for Domain-transfer Problem in the Context of Sentiment Analysis. In *CIKM*, pages 979–982, November 2007.

[128] Michael Thelen and Ellen Riloff. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 214–221, Philadelphia, Philadelphia, USA, July 2002. Association for Computational Linguistics.

[129] Matt Thomas, Bo Pang, and Lillian Lee. Get out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts. In *Proceedings of the*

5th Conference on Empirical Methods in Natural Language Processing (EMNLP-06), 2006.

[130] Peter Turney. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02)*, pages 417–421, Philadelphia, Philadelphia, USA, July 2002. Association for Computational Linguistics.

[131] Peter Turney and Michael Littman. Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada, 2002.

[132] Peter Turney and Michael Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21:315–346, 2003.

[133] Boris A. Uspensky. *A Poetics of Composition.* University of California Press, Berkeley, CA, 1973.

[134] Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. Developing Affective Lexical Resources. *PsychNology Journal*, 2(1):61–83, 2004.

[135] Cynthia M. Whissell. *The Dictionary of Affect in Language*, pages 113–131. Academic Press, New York, 1989.

[136] Cyntia M. Whissell and Kerry Charuk. A Dictionary of Affect in Language: II. Word Inclusion and Additional Validation. *Perceptual and Motor Skills*, 61(1):65–66, 1985.

[137] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of CIKM-05, the 14th ACM SIGIR Conference on Information and Knowledge Management*, Bremen, Germany, 2005.

[138] Janyce Wiebe. Tracking Point of View in Narrative. *Computational Linguistics*, 20(2):233–287, 1994.

[139] Janyce Wiebe. Instructions for Annotating Opinions in Newspaper Articles. Technical Report TR-01-101, University of Pittsburgh, Department of Computer Science, Pittsburgh, PA, 2002.

[140] Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane J. Litman, David R. Pierce, Ellen Riloff, Theresa Wilson, David S. Day, and

Mark T. Maybury. Recognizing and organizing opinions expressed in the world press. In *New Directions in Question Answering*, pages 12–19, Stanford, California, USA, March 2003.

[141] Janyce Wiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A Corpus Study of Evaluative and Speculative Language. In *Proceedings of the 2nd ACL SIGDial Workshop on Discourse and Dialogue*, Aalberg, Denmark, September 2001. Association for Computational Linguistics.

[142] Janyce Wiebe and Rada Mihalcea. Word Sense and Subjectivity. In *Proceedings of Association for Computational Linguistics (COLING/ACL-2006)*, pages 1065–1072, Sydney, Australia, July 2006. Association for Computational Linguistics.

[143] Janyce Wiebe and Ellen Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics*, volume 3406 of *Lecture Notes in Computer Science*, pages 475–486, Mexico City, Mexico, February 2005. Springer.

[144] Janyce Wiebe, Theresa Wilson, and Matthew Bell. Identifying Collocations for Recognizing Opinions. In *Proceedings of the ACL-01 Workshop on Collocation)*, Toulouse, France, July 2001. Association for Computational Linguistics.

[145] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning Subjective Language. *Computational Linguistics*, 30(3):277–308, 2004.

[146] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2–3):165–210, 2005.

[147] Janyce M. Wiebe. Learning Subjective Adjectives from Corpora. In *Proceedings of AAAI-00, the 17th National Conference on Artificial Intelligence*, pages 735–740, Menlo Park, CA, 2000. AAAI, AAAI Press / The MIT Press.

[148] Janyce M. Wiebe and Rebecca F. Bruce. Probabilistic Classifiers for Tracking Point of View. In *Proceedings of the Symposium on Empirical Methods in Discourse Interpretation and Generation, AAAI 1995 Spring Symposium Series*, pages 181–187, 1995.

[149] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and Use of a Gold-Standard Data Set for Subjectivity Classifications . In *Proceedings of the 37th*

Annual Meeting of the Association for Computational Linguistics (ACL-99), pages 246–253. Association for Computational Linguistics, 1999.

[150] Theresa Wilson and Janyce Wiebe. Annotating Opinions in the World Press. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*, pages 246–253, Sapporo, Japan, July 2003.

[151] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Recognizing Strong and Weak Opinion Clauses. *Computational Intelligence*, 2(22):73–99, 2006.

[152] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers, San Francisco, California, USA, 2005.

[153] Jianxin Yao, Gengfeng Wu, Jian Liu, and Yu Zheng. Using Bilingual Lexicon to Judge Sentiment Orientation of Chinese Words. In *Proceedings of the Sixth IEEE International Conference on Computer and Information Technology (CIT'06)*, page 25, Bhubaneswar, India, December 2006. IEEE Computer Society.

[154] D. Yarowsky. One Sense per Collocation. In *ARPA Workshop on Human Language Technology*, pages 266–271, 1993.

[155] Qiang Ye, Wen Shi, and Yijun Li. Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach. In *Proceedings of the Hawaii International Conference on System Sciences - HICSS. Track 3*, volume 3, page 53b, Kauai, Hawaii, USA, January 2006.

[156] Hong Yu and Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In Michael Collins and Mark Steedman, editors, *Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, Japan, 2003.

[157] Lotfy A. Zadeh. Calculus of Fuzzy Restrictions. In L.A. Zadeh, K.-S. Fu, K. Tanaka, and M. Shimura, editors, *Fuzzy Sets and their Applications to Cognitive and Decision Processes*, pages 1–40. Academic Press Inc., New-York, 1975.

[158] Lotfy A. Zadeh. PRUF — a Meaning Representation Language for Natural Languages. In R.R. Yager, S. Ovchinnikov, R.M. Tong, and H.T. Nguyen, editors, *Fuzzy Sets and Applications: Selected Papers by L.A. Zadeh*, pages 499–568. John Wiley & Sons, 1987.

[159] Taras Zagibalov and John Carroll. Unsupervised Classification of Sentiment and Objectivity in Chinese Text. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India, 2008.

[160] Zhu Zhang and Balaji Varadarajan. Utility Scoring of Product Reviews. In *Proceedings of the CIKM 2006: 15th ACM SIGIR International Conference on Information and Knowledge Management*, Washington DC, 2006.