An Evaluation Framework and Adaptive Architecture for Automated Sentiment Detection

Stefan Gindl, Johannes Liegl, Arno Scharl, and Albert Weichselbraun

Abstract. Analysts are often interested in how sentiment towards an organization, a product or a particular technology changes over time. Popular methods that process unstructured textual material to automatically detect sentiment based on tagged dictionaries are not capable of fulfilling this task, even when coupled with part-of-speech tagging, a standard component of most text processing toolkits that distinguishes grammatical categories such as article, noun, verb, and adverb. Small corpus size, ambiguity and subtle incremental change of tonal expressions between different versions of a document complicate sentiment detection. Parsing grammatical structures, by contrast, outperforms dictionary-based approaches in terms of reliability, but usually suffers from poor scalability due to its computational complexity. This work provides an overview of different dictionary- and machine-learning-based sentiment detection methods and evaluates them on several Web corpora. After identifying the shortcomings of these methods, the paper proposes an approach based on automatically building Tagged Linguistic Unit (TLU) databases to overcome the restrictions of dictionaries with a limited set of tagged tokens.

1 Introduction

Sentiment Detection (SD) is the part of Natural Language Processing (NLP) that deals with the automated extraction of opinions (the 'sentiment') out of unstructured text. The goal is to automatically decide whether an author expresses positive or negative sentiment towards a certain topic. The appeal of this research area lies in

Stefan Gindl, Johannes Liegl, and Arno Scharl
MODUL University Vienna, Department of New Media Technology, Austria
e-mail: `stefan.gindl,johannes.liegl,arno.scharl@modul.ac.at`

Albert Weichselbraun
Vienna University of Economics and Business Administration,
Research Institute for Computational Methods, Austria
e-mail: `albert.weichselbraun@wu-wien.ac.at`

S. Schaffert et al. (Eds.): Networked Knowledge - Networked Media, SCI 221, pp. 217–234.
springerlink.com

its wide range of possible applications, since reliable automated SD methods allow the analysis of large texts corpora beyond the limits of manual approaches.

The information obtained by this process may be used for several purposes. Applications include monitoring the launch and performance of commercial products, analyzing the electoral behavior of the public to guide political campaigns, or refining search engines to consider opinions. Yet, SD is a very ambitious problem to solve. NLP is one of the most challenging research areas in computer science, since natural languages are not as restrictive as formal languages. Natural language allows authors to express concepts in many different ways, which complicates automated analyses. Consider the following sentence:

> *The plot of the movie was banal and the actors were really clumsy.*

This sentence expresses a viewer's displeasure with a particular movie. Now consider the same sentence in the following context:

> *The plot of the movie was banal and the actors were really clumsy. However, I enjoyed it more than any other movie I have seen in the last few months!*

In this example, both sentences describe the same item (a particular movie), but differ in regard to the expressed sentiment. For an automated system, such constructs are very hard to evaluate. A human reader, by contrast, easily recognizes that the viewer liked the movie. Linguistic notions such as sarcasm or irony are even harder to spot by an algorithm.

This paper evaluates and compares several well-known SD techniques, such as the bag-of-words approach and maximum entropy modeling. Based on this analysis, we develop an alternative approach based on Tagged Linguistic Units (TLUs), annotating tokens and phrases with additional features such as part-of-speech (POS), context and topic.

The remainder of this article is structured as follows. Section 2 provides an overview of related work. Section 3 compares deep parsing strategies with approaches focusing on lexis. Section 4 describes state-of-the-art SD methods in greater detail, which are then evaluated in Section 5. After discussing the results, we identify weaknesses in current approaches and propose a novel method based on Tagged Linguistic Units in Section 6. Section 7 concludes the paper and presents an outlook on further research.

2 Related Work

The field of SD reveals emotional aspects of a written text, hinting at the opinion and intention of the author. This information can be used for several reasons: search engines can augment their results, marketing managers can find out why their product failed in a certain market, and political analysts can predict electoral behavior. The challenge of detecting sentiment in unstructured text leads to a vast amount of different approaches to tackle this task. Some of these only use binary decisions (a positive or negative sentiment), others use more sophisticated classifications.

The context of a sentiment term influences its meaning - e.g., in ’the president of the National Environment Trust’, the term ’trust’ refers to a large enterprise and not to ’confidence’. Wilson et al. [24] acknowledge the importance of context information by using a set of 28 features such as modifiers or adjacent terms, which are input to the AdaBoost machine learning approach.

Lexical units can also be distinguished from each other by using so called ‘appraisal taxonomies’ [22]. These contain information on the ’attitude’ (e.g., ‘appreciation’ or ‘affect’), the ‘orientation’ (positive vs. negative), the ‘force’ (can be increased by modifiers like ‘very’), or the ‘polarity’ (a binary decision depending on the existence of a negation trigger) of words.

Hatzivassiloglou and McKeown [4] base predicting the sentiment of adjectives on the hypothesis that conjoined adjectives may carry the same sentiment charge. Based on this hypothesis, their proposed system assigns an adjective with unknown sentiment the same sentiment value as its conjoined adjective.

Pang et al. [12] apply machine learning methods (Naive Bayes, Maximum Entropy Model, Support Vector Machines) in combination with a bag-of-features (i.e., a collection of terms with certain characteristics such as a sentiment) framework to a data set containing reviews from the ‘Internet Movie Database’. Pang and Lee present a refinement of this approach in their later work [11], where they involve a previous subjectivity classification (i.e., a method capable of discriminating sentences into subjective and objective ones). As compared to objective sentences that are only used to describe facts, subjective sentences are supposed to reflect the opinion an author intends to express. Kushal et al. [7] also apply three machine learning methods to product reviews, comparing their results to a simple baseline algorithm. Mullen and Collier [10] work with Support Vector Machines, where a list of terms and their sentiment values (i.e., a value corresponding to the general affinity of the term to express positive or negative opinion) represents the features. A generic process using Pointwise Mutual Information then determines the sentiment values of these terms.

Yu and Hatzivassiloglou [25] present an approach for subjectivity classification using a Naive Bayes classifier. Riloff and Wiebe [14] present a bootstrapping approach to automatically create large training sets in order to learn extraction patterns for subjectivity. In another work, Wiebe and Riloff [23] produce training data for the training of a Naive Bayes subjectivity classifier by employing a rule-based classifier. Subasic and Huettner [19] apply fuzzy methods to analysing affect in writings. Blitzer et al. [2] present an approach using similarities between differing domains in order to adapt a sentiment classifier to a new domain. Ding et al. [3] determine the sentiment of a sentence in regard to a specific object within this sentence (in this case, objects refer to products like cameras). Conjunction rules help accomplish this task for both the usage within a sentence as well as multiple sentences. Another feature is a distance function, which determines the correlation of sentiment terms considering their absolute distance to a specific object.

3 Lexical Approaches versus Full Parsing

Capturing the evolution of information spaces calls for a new generation of robust, language-independent and distributed natural language processing techniques optimized for throughput and scalability. From a stakeholder perspective, sentiment expressed in textual material (e.g., news media coverage) is of particular interest [17]. Automated methods to compute sentiment, however, usually belong to one of the following two categories: (i) low-overhead approaches that focus on the lexis of text, and (ii) full parsing of grammatical structures, which improves the accuracy of results but suffers from poor scalability. This paper presents a new method that falls into the first category but aims to improve the quality of results by building an adaptive databases of *tagged linguistic units*. Such a database helps ensure scalability, preserve context information and process heterogeneous data sources.

Most research projects that apply automated sentiment detection techniques such as the *US Election 2008 Web Monitor* (www.ecoresearch.net/election2008) or the *Media Watch on Climate Change* (www.ecoresearch.net/climate) typically gather a large corpus of text compiled from many sources and sampled in regular intervals. Using POS tagged and partially parsed corpora to identify relevant sketches (= co-occurrence lists for grammatical patterns provided by a grammar rule engine) improves the performance of existing SD-techniques [5, 6], but processing arbitrarily long blocks of text still requires a fundamentally new strategy. The ability to work with very short textual segments is paramount when trying to analyze the *evolution* of knowledge reflected in corpora. Longitudinal studies of specific topics or events often yield few additional occurrences of a term in a given interval, as incremental changes to existing documents are common. This complicates the analysis, because the validity of many text processing methods depends on corpus size and frequency of target terms.

Given the unresolved scalability issues of SD methods that rely on full parsing, this paper describes attempts to extend and improve lexis-based approaches with a special focus on context-aware processing. The next section will summarize standard dictionary-based SD methods and compare them to machine learning approaches.

4 Algorithm Description

This section focuses on the most common SD methods (arithmetic, machine learning based and combined), and describes a framework for evaluating them based on three different corpora compiled from Web resources available to the public:

- *Amazon* (www.amazon.com) provides customer reviews ranging from "one-star" (low recommendation) to "five-star" (high recommendation) ratings. The Amazon data set consists of 165,746 book reviews and contains 1,539,058 sentences.

- The *Internet Movie Database (IMDb)* (www.imdb.com) contains 2000 reviews comprising 69,207 sentences. The IMDb data set was also used in [12], thus the reviews already carry information on positive and negative sentiment.
- *TripAdvisor* (www.tripadvisor.com) provides reviews of holiday destinations. It contains 7554 reviews with ratings from one to five stars, where one star indicates a very low recommendation and five stars a high recommendation. This data set comprises 62,818 sentences.

Amazon and TripAdvisor rate each review on a scale from one to five stars. We generalize these ratings and consider all reviews with a rating lower than three as negative, all reviews with a rating greater than three as positive, and ratings of three as neutral. In order to avoid adulterated results, we use balanced versions of the data sets - i.e., subsets of the original data containing exactly the same number of positive and negative reviews. The Amazon data set contains 21,458 negative, 130,061 positive and 14,227 neutral reviews. The TripAdvisor data set consists of 1105 negative, 5673 positive and 776 neutral reviews. The balancing filter yields a total of 420,840 sentences from Amazon and 17,768 sentences for TripAdvisor (IMDb provides an already balanced data set).

4.1 Arithmetic Methods

The arithmetic methods are based on tagged dictionaries, which contain sentiment terms with corresponding sentiment values in a closed interval [-1,1]. For example, ‘champion’ is a positive word carrying the sentiment value ‘1’, whereas ‘charlatan’ carries the negative value ‘-1’. The dictionary contains a total of 8267 sentiment terms, 5072 of them positive and 3195 negative. The tagged dictionary is not domain-specific, which helps draw conclusions on its general applicability. Subjecting the General Inquirer (www.wjh.harvard.edu/˜inquirer/) dictionary to a reverse lemmatization process yielded 7302 terms, 965 additional entries were manually retrieved from a sample of online blogs. Arithmetic algorithms browse through the reviews and search for terms contained in the dictionary. The number of detected terms gives information about the overall sentiment value of a sentence. Each of the following methods calculates the overall sentiment of a review by summing up all sentiment values of the individual sentences.

- *Simple SD (SSD)* counts the values of sentiment terms in a sentence. If the sum of these values is positive (negative), the sentence is considered to have a positive (negative) sentiment. If a negation trigger such as ‘not’ and ‘never’ occurs directly before a sentiment term (e.g., ‘The proposal was not approved’), the value of this term is multiplied by ‘-1’, resulting in an inverted sentiment value. We used this method as a simple baseline approach towards SD.
- *Extended SD (ESD)*. This method incorporates other semantic components affecting the results. We extended the former detection method by so called modifier terms (e.g., ‘very’, ‘rather’). If such a term occurs before a sentiment word, the value of the sentiment word is either increased or decreased, depending on the orientation of the modifier. The term ‘very’ increases a sentiment value (e.g.,

'The candidate is very charming.'); we, therefore, multiply the original value by '1.5'. In the case of the decreasing term 'little', we multiply the term's sentiment by '0.5' (e.g., 'The patient felt little pain.').

- *Adjective Detection (AD).* Adjectives are often used to express sentiment. For that reason, we investigated the outcome of a SD method using only adjectives. In order to limit the method to adjectives, we applied the POS tagger of the OpenNLP project (opennlp.sourceforge.net).
- *Detailed Part-Of-Speech Detection (DetPOSD).*This method applies POS tagging to determine the scope of a negation trigger. For each occurrence of a term with a semantic value, the method tries to identify a negation trigger that instructs the algorithm to multiply the sentiment value by '-1'. Certain constituents help refine this procedure and avoid negation triggers from impacting the complete sentence (although they were not meant to do so). If a noun phrase respectively a verb phrase is positioned between the sentiment word and the negation trigger, this term is regarded as negated and the original sentiment word will remain unaffected. Figure 4.1 shows an illustration of this procedure: a negation trigger occurs at the beginning of the sentence (NT) and a sentiment token at the end (SentT). Between these are placed a number of arbitrary tokens (AT; this can be determiners, adjectives etc.) that do not influence the negation. Yet, the stop token (ST; this can be either verbs or nouns) decides that the trigger does not influence the sentiment token, and thus, the sentiment token remains as being not negated.

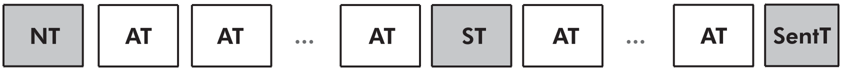

Fig. 1 Scope determination for negation triggers (NT=Negation Trigger, AT=Arbitrary Token, ST=Stop Token, SentT=Sentiment Token)

4.2 Machine Learning Methods

In the following, we compare three different methods: a language model, a Naive Bayes classifier and a Maximum Entropy Model (Section 5). These methods do not use a tagged dictionary but build their knowledge base in a training step. The existing classification of the data sets suggests using supervised learning. Performing experiments with a generic training set and evaluating the results on a domain-specific test set sheds light on the methods' universality.

The experiments on *generic knowledge bases* trained the Language Model and Naive Bayes algorithm on the IMDb data set, using the TripAdvisor and Amazon sets for testing purposes. In a follow-up step, a model on the TripAdvisor data set was trained to be tested on the IMDb set. While this procedure allows training and testing on the complete datasets (avoiding the need to split into training and test sets), it faces domain-dependent constraints (since machine learning methods tend to strongly fit to the domain they were trained on). Training the Maximum Entropy Model with a part of each data set as a training set and the other part as a test set yields the *domain specific knowledge base*.

The LingPipe libraries[1] and OpenNLP MaxEnt Package[2] helped streamline the implementation of the different learning methods. The following itemization provides a detailed explanation of the three methods.

- *Language Model (LM).* The evaluation uses an implementation of a LingPipe language model classifier to create a language model. A language model is a probabilistic representation of a sequence of words. We trained the language model by separately providing the classifier with positive and negative reviews (thus, the classifier *knew* the sentiment class of the presented review). In the next step, the created model had to predict the sentiment class of reviews of unknown sentiment.
- *Naive Bayes (NB).* A Naive Bayes classifier proceeds on the conditional independence assumption, which expects the attributes allowing a classification to be independent from each other [11]. Although most real-world applications violate this assumption, the algorithm yields surprisingly good results. Zhang [26] explains this good performance by suggesting that two attributes may depend on each other in a given data set, but the dependence may be distributed evenly in each class.
- *Maximum Entropy Model (MaxEnt).* Maximum Entropy Models can integrate features from heterogeneous information sources without posing strong independence assumptions like the Naive Bayes approach. Features correspond to constraints in the model, and the Generalized Iterative Scaling algorithm [13] outputs the model, which maximizes the entropy among the constraints. The method yields the model preserving the most uncertainty. This is desired, because every other model would add information that is not justified by empirical evidence (i.e. the training data) [1, 9]. Each data set required a unique Maximum Entropy Model, using only unigrams as features given their good performance in previous studies [12]. One-third of the reviews of the corresponding dataset were used to train the model, leaving the remaining two thirds of the dataset for evaluation purposes.

Generic approaches to sentiment detection represent a challenging problem. Domain-specific methods are generally assumed to deliver superior results. The evaluation of NB and LM across domains allowed investigating whether these methods could be used for multiple domains without having access to domain-specific training corpora. Alternatively, the Maximum Entropy Model was trained on a subset of a corpus and tested on another subset of the same corpus, yielding a model specifically fitted to the domain.

[1] http://alias-i.com/lingpipe/

[2] http://maxent.sourceforge.net/

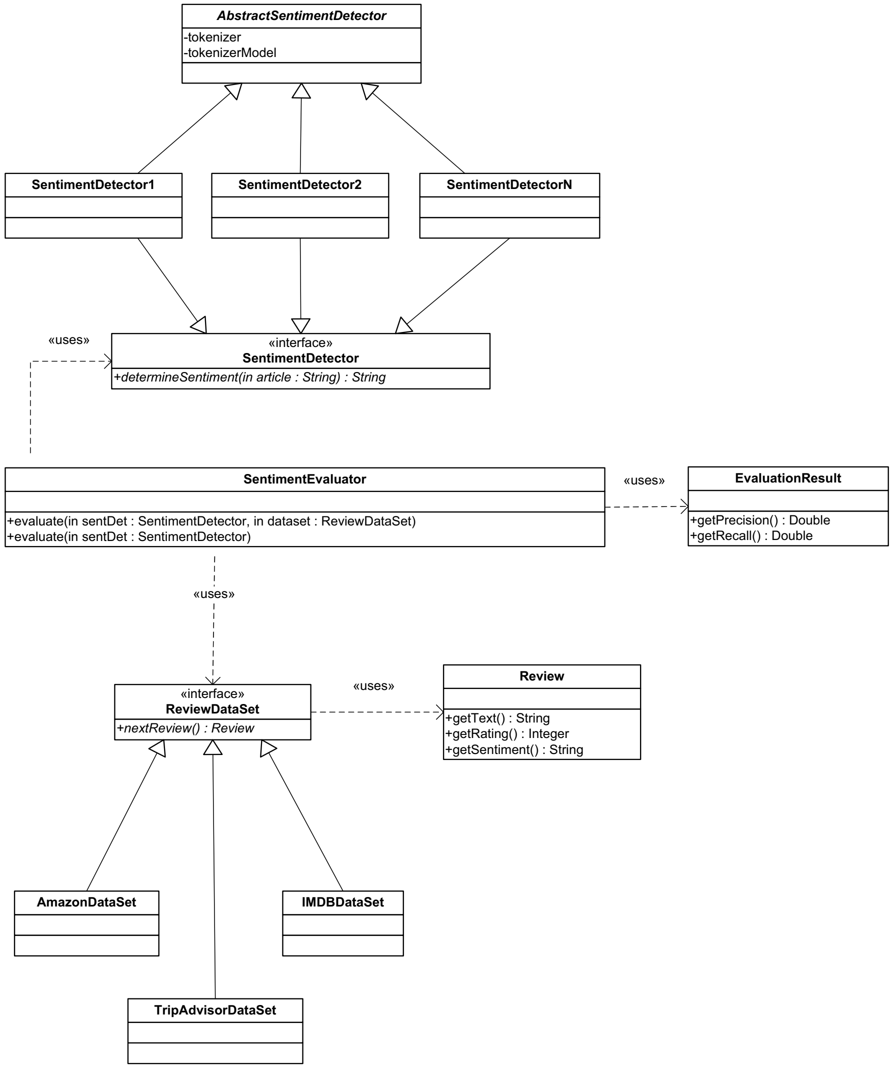

Fig. 2 UML diagram of the evaluation framework

4.3 System Architecture

The evaluation framework presented in the following allows comparing the results of the various SD methods through the SentimentDetector interface. New data is integrated by implementing the ReviewDataSet interface. To evaluate an SD method on a data set, implementations of both have to be passed to the evaluation method of the SentimentEvaluator. This component iterates the reviews in the specified data set and applies the specified SentimentDetector on each of them. The results of the SD for each review is then compared with the review's original rating and the outcome is classified as true positive, true negative, false positive or false negative for the calculation of statistical measures (see Subsection 5.2 for details on the used statistical measures). Figure 2 shows a UML diagram of the evaluation framework.

5 Evaluation

The evaluation focuses on the SD method's ability to put a review into the right polarity class (positive or negative). The IMDb data set is already divided into positive and negative classes. For the Amazon and TripAdvisor data sets, a pre-processing module maps user ratings between 'one star' and 'five stars' to the classes 'negative', 'neutral', and 'positive' as outlined in Section 4.

5.1 Statistical Properties of the Corpora and Implications

This section describes the statistical structure of the evaluation corpora. Table 1 lists the minimum, maximum, average and standard deviation of (a) the number of occurring positive and negative tokens from the tagged dictionary in the data set, (b) the absolute number of tokens, and (c) the number of sentences in a review.

The results of the descriptive statistics show that the Amazon and TripAdvisor data sets are very heterogeneous, caused by a number of extreme outliers (according to their large standard deviation, which even exceeds the average in the Amazon data set for the positive and negative tokens). The IMDb data, by contrast, presents itself as being more homogeneous. Another advantage of this set is the fact that each

Table 1 Statistical characteristics of the review data sets

Data set	Param.	Pos. Tokens	Neg. Tokens	Single Terms	Sent.
IMDb	Max	164	138	2753	124
	Min	1	1	18	1
	Avg	45.76	37.19	761.56	34.6
	StdDev	21.75	18.18	334.82	16.15
TripAdvisor	Max	70	43	1240	62
	Min	0	0	1	1
	Avg	8.66	4.81	160.11	8.04
	StdDev	7.71	5.05	130.23	5.83
Amazon	Max	260	226	4878	157
	Min	0	0	1	1
	Avg	12.32	7.86	211.91	9.81
	StdDev	12.81	9.36	209.73	8.99

sentence contains at least one sentiment token of the positive and negative class. In the case of the other two sets, this is not ensured. A number of zero sentiment tokens in a review would lead to a result of zero for the review, which is then considered as being a neutral review. We do not filter such reviews, since we assume that reviews containing no sentiment token express a neutral opinion.

5.2 Detailed Results

The evaluation considers five statistical parameters: recall, precision, accuracy, F measure, and Cohen's kappa coefficient. Recall is a measure for the completeness of a detection method - i.e., it shows how many of the requested objects could actually be found. On the other hand, precision provides a measure for the number of objects that have been identified correctly. The accuracy is the ratio of all correctly identified objects and the (either correctly or incorrectly) classified objects. The F measure combines recall and precision. Cohen's kappa coefficient is normally used to measure the inter-rater-reliability, that is, how strongly different raters classifying a number of objects agree on the classification of these objects.

We calculate recall and precision for the positive and negative class separately. Separate precision and recall results for the positive and negative class are required because the classifiers we use can also return a neutral result (namely when no sentiment token occurs in a review). Therefore, a document that is not negative is not automatically positive. This procedure also leads to separate results for Cohen's kappa value as well as the F measure. Tables 2 to 4 show the detailed evaluation results of each data set.

Using the IMDb data set resulted in fairly balanced results. The arithmetic methods achieve good results for the detection of positive as well as negative reviews

Table 2 Evaluation of the SD methods applied to the IMDb data set (FM=F measure)

	IMDb Data Set									
	Positive Sentiment					**Negative Sentiment**				
Detection Method	**Rec.**	**Prec.**	**Acc.**	**Cohen's Kappa**	**FM**	**Rec.**	**Prec.**	**Acc.**	**Cohen's Kappa**	**FM**
	Generic Methods									
SSD	70.1	63.38	64.8	0.3	0.67	59.4	66.52	64.75	0.29	0.63
ESD	68.2	64.1	65	0.3	0.66	61.9	66.06	65.05	0.3	0.64
AD	67.7	60.39	61.65	0.23	0.64	55.4	63.17	61.55	0.23	0.59
DetPOSD	69.3	63.52	64.75	0.29	0.66	60.3	66.26	64.8	0.3	0.63
LM	16.1	70.61	54.7	0.09	0.26	**90.7**	53.2	55.45	0.11	0.67
NB	**96.5**	51.09	52.05	0.04	0.67	5.4	65.85	51.3	0.03	0.1
	Domain-Specific Method									
MaxEnt	80.51	**83.39**	**82.23**	**0.64**	**0.82**	83.96	**81.16**	**82.23**	**0.64**	**0.83**

Table 3 Evaluation of the SD methods applied to the TripAdvisor data set (FM=F Measure)

	TripAdvisor Data Set									
	Positive Sentiment					**Negative Sentiment**				
Detection Method	**Rec.**	**Prec.**	**Acc.**	**Cohen's Kappa**	**FM**	**Rec.**	**Prec.**	**Acc.**	**Cohen's Kappa**	**FM**
	Generic Methods									
SSD	93.39	65.61	72.22	0.44	0.77	30.41	86.6	62.85	0.26	0.45
ESD	92.76	66.13	72.62	0.45	0.77	32.04	85.71	63.35	0.27	0.47
AD	83.62	65.67	69.95	0.4	0.74	26.52	76.1	59.1	0.18	0.39
DetPOSD	93.21	66.15	72.76	0.46	0.77	31.67	86.63	63.39	0.27	0.46
LM	40.09	57.53	55.25	0.1	0.47	63.44	58.91	59.59	0.19	0.61
NB	49.77	**76.92**	67.42	0.35	0.6	**77.01**	67.86	70.27	0.41	**0.72**
	Domain-Specific Method									
MaxEnt	**93.89**	67.71	**74.56**	**0.49**	**0.79**	55.22	**90.04**	**74.56**	**0.49**	0.68

Table 4 Evaluation of the SD methods applied to the Amazon data set (FM=F Measure)

	Amazon Data Set									
	Positive Sentiment					**Negative Sentiment**				
Detection Method	**Rec.**	**Prec.**	**Acc.**	**Cohen's Kappa**	**FM**	**Rec.**	**Prec.**	**Acc.**	**Cohen's Kappa**	**FM**
	Generic Methods									
SSD	77.23	57.27	59.8	0.2	0.66	39.78	66.96	60.08	0.2	0.5
ESD	75.88	57.56	59.97	0.2	0.65	41.5	66.39	60.25	0.2	0.51
AD	59.73	55.88	56.28	0.13	0.58	39.51	63.06	58.18	0.16	0.49
DetPOSD	76.89	57.59	60.13	0.26	0.66	40.79	67.11	60.4	0.21	0.51
LM	41.67	72.9	63.09	0.26	0.53	75.17	62	64.55	0.29	0.68
NB	47.98	68.88	63.15	0.26	0.57	69.86	63.14	64.54	0.29	0.66
	Domain-Specific Method									
MaxEnt	**78.73**	**87.02**	**83.49**	**0.67**	**0.83**	**88.26**	**80.58**	**83.49**	**0.67**	**0.84**

(considering the usage of a domain-independent tagged dictionary). The TripAdvisor data set yields the best results for the detection of positive reviews. Yet, this outstanding performance is accompanied by quite poor results in the detection of negative sentences. The Amazon data set also satisfyingly identifies positive reviews at the cost of an inferior precision for negative reviews.

The better results in the positive category represents a surprising result, since the tagged dictionary contains more negative than positive sentiment tokens (5072 negative in contrast to 3195 positive ones). In spite of this, the statistical analysis in 5.1 shows that in all data sets, a larger number of positive tokens occurred. This fact leads to the assumption that customers use positive tokens more frequently than negative ones and that positive words might also be used in order to express a negative opinion towards a movie, book or holiday destination (e.g., in the case of humor or sarcasm). Additional context information would help resolve some of these cases and determine sentiment more accurately.

The Maximum Entropy Model, which entails domain-specific knowledge, outperforms the other methods. It produces results with the highest precision, recall and kappa value in the negative classification task as well as in the positive. These findings do not suggest that arithmetic SD generally provides inferior results, but that the knowledge base and the application domain play an important role in the identification of negative sentiment. Methods that consider the domain context (see Section 6) therefore have the potential to yield much better results.

5.3 Discussion

The evaluation results show that the presented SD methods have their strength in the identification of reviews with positive sentiment (high recall). Only a relatively small number was overlooked by the algorithms. On the other hand, the method's precision is less satisfactory. A rather high amount of items has been incorrectly identified as having a positive sentiment.

On the TripAdvisor data set, the methods achieve an excellent recall between 83% and 93% without any decrease in precision. We assume that the writing style of this kind of data alleviates the SD - at least for positive sentiment. The results for the detection of negative sentiment are less encouraging. It seems difficult to correctly extract reviews with negative sentiment (very low recall). Yet, precision does not decrease to the same extent. As for precision, the SD on the TripAdvisor data set again outperforms the results obtained with the other data set.

We assume that the structure of the reviews in the IMDb and Amazon data set strongly influences the outcomes. Reviews of movies and books often integrate plot summaries into the evaluation. In the case of love films, for example, a notable number of words carrying positive sentiment like 'love', or 'happy' (if the film has a happy end) will occur, even when the reviewer dislikes the product. The same consideration applies to horror films or thrillers that contain negative vocabulary in the plot summary.

The Maximum Entropy model clearly outperformed the other SD methods, particularly in detecting negative sentiment, which is not surprising given that it has been trained and tested within the same domain. This should guide future research and favors domain-specific components whenever the required context information is available. Building on this insight, the following section proposes to build databases of Tagged Linguistic Units (TLU). Such a repository contains a comprehensive list of terms of a certain language together with their significance for emotional speech (i.e., their sentiment value) and additional metadata.

6 Tagged Linguistic Units

Tagged Linguistic Units (TLUs) comprise units of linguistic content such as terms and phrases, coupled with a set of annotations (e.g., POS tags, topic or prevalent context). They combine the advantages of methods that go beyond lexis without inheriting the full complexity of grammar parsing. The following sections outline the generation of TLU databases and their application to sentiment detection.

6.1 Database Creation

As already mentioned in section 4, simple SD methods that do not use machine learning algorithms on narrowly defined domains rely on a tagged dictionary that distinguishes between positive- and negative-valued sentiment words [16]. Such dictionaries typically contain a few thousand mappings from words to their associated sentiment values - e.g., the General Inquirer [18]). They can be subjected to a reverse lemmatization procedure, adding inflections to the initial list of sentiment words. Even assuming such an extended tagged dictionary, dictionary-based approaches do not take the context of sentiment words into account, which limits their usefulness in corporate knowledge architectures.

The rest of this paper addresses this shortcoming by proposing a hybrid method based on spreading activation networks coupled with machine learning algorithms for assigning sentiment values to linguistic units. For this purpose, the following linguistic units for computing sentiment will be distinguished: *unigram* (single word), *n-gram* (multiple-word units of meaning), and *concepts* (units of meaning not tied to a particular lexical form and represented via rules or regular expressions, e.g. *climate change* $\Leftrightarrow$ *global warming*).

A sentiment value and a context (e.g., part-of-speech, geographic location and named entity) are assigned to each linguistic unit. For a given amount of text, these mappings taken together are the building blocks of a *Tagged Linguistic Unit (TLU)* database. The sentiment values stored in this database are constantly being updated based on new data from the knowledge acquisition services and can be customized for specific domains, applications or users. Generating and using a TLU database instead of a tagged dictionary that only contains words and binary classifications allows a fine-grained differentiation between sentiment values associated with morphologically similar but semantically different linguistic units such as *cell*, *fuel cell* and *prison cell* through the consideration of contextual information like POS tags, geo tags and named entity tags.

Work by Scharl et al. [15] has demonstrated the usefulness of assigning sentiment values to geographic locations and also shows how heavily these values depend on other context dimensions. Future research will address these dependencies by combining tags with more sophisticated context information as for instance hierarchical classifications [20] or topic tags. This approach (i) is language-independent in the sense that only a small set of seed terms (e.g., 100 positive and 100 negative terms) and grammar patterns would be required to initialize the machine learning algorithm

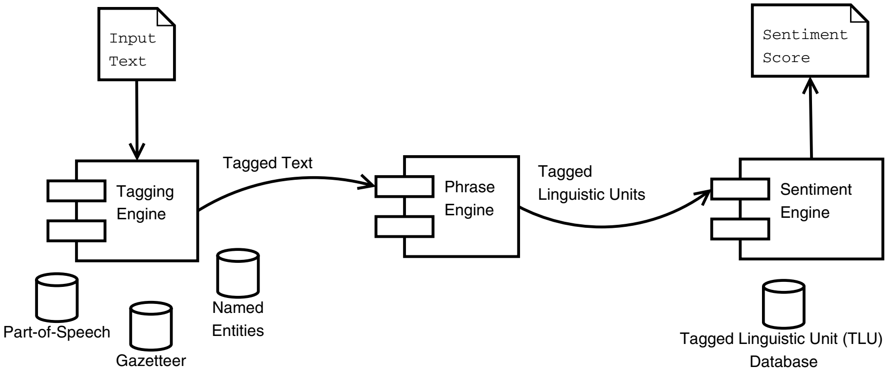

Fig. 3 Sentiment scoring based on linguistic units

and fine-tune sentiment values to any language that is decomposable into unigrams, n-grams and concepts, (ii) is not restricted to the sentiment categories 'positive' and 'negative', but supports an arbitrary number of linguistic categorizations such as weak ⟷ strong, passive ⟷ active, etc., (iii) ensures that every sentence or document can be annotated; traditional approaches often encounter sentences that do not contain any of the words listed in the tagged dictionary.

Figure 3 illustrates sentiment scoring based on linguistic units. The phrase engine identifies the linguistic units.

The tagging engine identifies part-of-speech tags, named entities, and geographic locations. The sentiment engine processes linguistic units and associated tags based on the data in the tagged linguistic units database, computing a sentiment value for the given text. Tagging provides important background information for these tasks. In the most straightforward case, the sentiment of linguistic units, as for instance the word `like`, depends on the assigned part-of-speech tag (`like/VB` versus `like/IN`). In more complex cases, named entity tags or even geo tags might be necessary to correctly identify the TLU's sentiment value (e.g., in the case of `National Environment Trust`).

6.2 *Iterative Extension and Optimization*

As outlined in the previous section, TLU databases can be easily customized to specific domains and use cases. A domain-specific corpus, language-specific grammar rules and a set of seed terms with "known" sentiment values (e.g., from conventional tagged dictionaries such as the General Inquirer repository) initialize the TLU database. The architecture identifies unknown linguistic units in the corpus and determines their sentiment value as illustrated in Figure 4.

The tagging component marks sentences with part-of-speech tags and identifies named entities such as people, organizations, and geographic locations. Combining co-occurrence analysis with a grammar rule engine yields candidate terms for

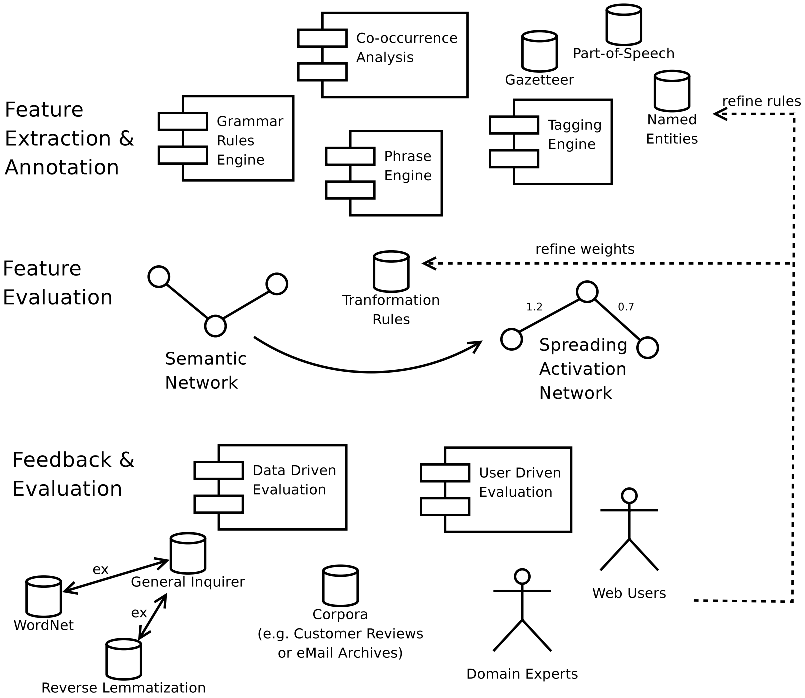

Fig. 4 Iterative fine-tuning of the Tagged Linguistic Unit (TLU) database

extending the TLU database. Annotating these terms with named entity tags and encoding characteristic grammatical patterns and known phrases creates a complex semantic network, which describes the relations between linguistic units.

Liu et al. [8] demonstrated how decomposing and translating semantic networks based on heuristic rules yield a spreading activation network for extending domain ontologies. Applying this approach to identifying and tracking tagged linguistic units builds a spreading activation network used to distribute the sentiment charges between the units based on the features and annotations generated during the annotation step. Activation of concepts with known sentiment charges in accordance to sign and strength of the charge leads to the propagation of energy pulses through the network, eventually distributing charges to all linguistic units. Analyzing the sentiment values' variance allows estimating confidence levels and identifying synonym ↔ antonym relationships.

Feedback gathered in the evaluation step adjusts and optimizes the transformation rules for a given domain and corpus, improving the quality of the TLU database with every subsequent step. Automatic data-driven evaluation on a TLU level will help assess overall performance. Using the evaluation framework outlined in Section 5 on various publicly available Web corpora will provide test cases for TLU-based sentiment detection. Automated methods will be complemented by user-driven evaluations from domain experts and Web users. The feedback gathered by the data- and user-driven evaluations will be utilized to refine the transformation rules of the feature evaluation, and to identify candidate patterns for the inclusion into the databases of the grammar rule engine and the phrase engine.

Automatically generating TLU databases faces the problem of determining the correct charge (+0.4 vs. -0.4, for example) of the sentiment value to be assigned to the linguistics unit. The problem arises from the fact that synonyms and antonyms have very similar (co-)occurrence patterns in a given corpus. Advanced relation discovery techniques developed within the AVALON project [21] will help overcome this challenge and facilitate the automation of this classification process. The machine learning algorithms will be trained and evaluated on augmented tagged dictionaries (created through reverse lemmatization and adding WordNet synonym and antonym pairs), as well as on public pre-tagged corpora.

7 Conclusion and Outlook

Simple approaches to sentiment detection based on patterns of co-occurrence with terms from tagged dictionaries scale well but provide less accurate results compared to complex methods that require a full parsing of sentence structures. The sheer volume of textual data and economic considerations, however, frequently rule out the most sophisticated approaches. Continuously updated databases of tagged linguistic units aim to balance accuracy and throughput. They add an adaptive layer to static sentiment detection approaches based on tagged dictionaries, which still tend to be compiled manually.

Preliminary results from the described approach are promising. Following a formal evaluation of different approaches to sentiment detection, recall and precision were significantly improved by adding WordNet synonyms and antonyms to the tagged dictionary (only considering synsets with high frequencies to exclude rare expressions). Currently, terms extracted from media corpora serve as candidates for assigning sentiment values via co-occurrence analysis, which will further extend the tagged dictionary.

Text mining projects have to process hundreds of thousands or millions of documents in short intervals. Thus they significantly benefit from accurate methods of determining sentiment with minimal computational requirements at run time. While the creation of tagged linguistic unit databases is computationally intense, the overhead of applying them within annotation components remains small. Improved sentiment detection algorithms will encourage their use in both academic and commercial applications. Refined versions of the sentiment detection methods presented in this paper will generate a richer set of context information (e.g., ontology concepts or explicit references to other types of structured knowledge), and consider this information in the scoring process.

Acknowledgment

This work was developed as a part of the RAVEN (Relation Analysis and Visualization for Evolving Networks; www.modul.ac.at/nmt/raven) research project funded by the Austrian Ministry of Transport, Innovation & Technology (BMVIT) and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT (www.fit-it.at). The authors wish to thank Arinya Eller for proof-reading the manuscript and Ivo Ponocny for his valuable feedback and suggestions.

References

1. Berger, A.L., Pietra, S.D., Pietra, V.J.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–71 (1996)
2. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 440–447 (June 2007)
3. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: WSDM 2008: Proceedings of the international conference on Web search and web data mining, Palo Alto, California, USA, pp. 231–240. ACM, New York (2008)
4. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Morristown, NJ, pp. 174–181. Association for Computational Linguistics (1997)
5. Kilgarriff, A., Evans, R., Koeling, R., Rundell, M., Tugwell, D.: Waspbench: A lexicographer's workbench supporting state-of-the-art word sense disambiguation. In: 10th Conference on European Chapter of the Association For Computational Linguistics, Morristown, USA, Association for Computational Linguistics (2003)
6. Kilgarriff, A., Rychl, P., Smrz, P., Tugwell, D.: The Sketch engine. In: 11th Euralex international Congress. Lorient, France (2004)
7. Kushal, D., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: WWW 2003: Proceedings of the twelfth international conference on World Wide Web, pp. 519–528. ACM Press, New York (2003)
8. Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. Journal of Universal Knowledge Management (1), 50–58 (2005), http://www.jukm.org/jukm_0_1/semi_automatic_ontology_extension
9. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge (1999)
10. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources (2004)
11. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts (September 2004)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP (2002)

13. Ratnaparkhi, A.: Maximum entropy models for natural language ambiguity resolution (1998)
14. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003) (2003)
15. Scharl, A., Dickinger, A., Weichselbraun, A.: Analyzing news media coverage to acquire and structure tourism knowledge. Information Technology and Tourism 10(1), 3–17 (2008)
16. Scharl, A., Pollach, I., Bauer, C.: Determining the semantic orientation of web-based corpora. In: Liu, J., Cheung, Y.-m., Yin, H. (eds.) IDEAL 2003. LNCS, vol. 2690, pp. 840–849. Springer, Heidelberg (2003)
17. Scharl, A., Weichselbraun, A.: An automated approach to investigating the online media coverage of us presidential elections. Journal of Information Technology & Politics 5(1), 121–132 (2008)
18. Stone, P.J.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
19. Subasic, P., Huettner, A.: Affect analysis of text using fuzzy semantic typing. IEEE Transaction on Fuzzy Systems 9(4), 483–496 (2001)
20. Weichselbraun, A.: Ontologiebasierende Textklassifikation mittels mathematischer Verfahren. PhD thesis, Vienna University of Economics and Business Administration (2004)
21. Weichselbraun, A., Wohlgenannt, G., Scharl, A., Granitzer, M., Neidhart, T., Juffinger, A.: Applying vector space models to ontology link type suggestion. In: 4th International Conference on Innovations in Information Technology, Dubai, United Arab Emirates, pp. 566–570. IEEE Computer Society Press, Los Alamitos (2007)
22. Whitelaw, C., Garg, N., Argamon, S.: Using Appraisal Taxonomies for Sentiment Analysis. In: Proceedings of MCLC 2005, the 2nd Midwest Computational Linguistic Colloquium, Columbus, US (2005)
23. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 486–497. Springer, Heidelberg (2005)
24. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA (2005)
25. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Collins, M., Steedman, M. (eds.) Proceedings of EMNLP 2003, 8th Conference on Empirical Methods in Natural Language Processing, Sapporo, JP, pp. 129–136 (2003)
26. Zhang, H.: The optimality of naive bayes. In: Barr, V., Markov, Z. (eds.) FLAIRS Conference. AAAI Press, Menlo Park (2004)