Using Games with a Purpose and Bootstrapping to Create Domain-Specific Sentiment Lexicons

Albert Weichselbraun
University of Applied Sciences
HTW Chur
Ringstraße 34
7000 Chur, Switzerland
albert.weichsel-braun@htwchur.ch

Stefan Gindl
MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1
1190 Vienna, Austria
stefan.gindl@modul.ac.at

Arno Scharl
MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1
1190 Vienna, Austria
arno.scharl@modul.ac.at

ABSTRACT

Sentiment detection analyzes the positive or negative polarity of text. The field has received considerable attention in recent years, since it plays an important role in providing means to assess user opinions regarding an organization's products, services, or actions.

Approaches towards sentiment detection include machine learning techniques as well as computationally less expensive methods. Both approaches rely on the use of language-specific sentiment lexicons, which are lists of sentiment terms with their corresponding sentiment value. The effort involved in creating, customizing, and extending sentiment lexicons is considerable, particularly if less common languages and domains are targeted without access to appropriate language resources.

This paper proposes a semi-automatic approach for the creation of sentiment lexicons which assigns sentiment values to sentiment terms via crowd-sourcing. Furthermore, it introduces a bootstrapping process operating on unlabeled domain documents to extend the created lexicons, and to customize them according to the particular use case. This process considers sentiment terms as well as sentiment indicators occurring in the discourse surrounding a particular topic. Such indicators are associated with a positive or negative context in a particular domain, but might have a neutral connotation in other domains.

A formal evaluation shows that bootstrapping considerably improves the method's recall. Automatically created lexicons yield a performance comparable to professionally created language resources such as the General Inquirer.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - Linguistic Processing;
H.5.3 [**Group and Organization Interfaces**]: Collaborative Computing

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

General Terms

Algorithms, Languages, Performance

Keywords

Sentiment Detection, Bootstrapping, Language Resources, Sentiment Lexicon, Crowd-Sourcing

1. INTRODUCTION

Sentiment detection has attracted a lot of research interest in recent years. With the emergence of freely available opinions on the Web the need for efficient methods to interpret these opinions has arisen. Automated sentiment detection is capable of accomplishing this task. It facilitates means of large-scale investigation previously unmanageable for humans, such as tracking political campaigns on the Web or market research in forums or blogs. Reliable sentiment detection is heavily dependent on the comprehensiveness and accuracy of the underlying a-priory knowledge, in most cases a so-called sentiment lexicon. This lexicon contains opinionated terms and is usually manually compiled. The occurrence of these terms in a document serves as indicator for "positiveness" or "negativeness" of a document. Manually compiling sentiment lexicons can be cumbersome and such lexicons may lack comprehensiveness, especially in the case of less-spoken languages. The presented method combines a crowd-sourcing technique, which is used for creating an initial sentiment lexicon, with a bootstrapping approach that automatically expands sentiment lexicons with additional terms. As input serves an unlabeled text corpus, from which a labeled corpus is iteratively extracted. Based on this labeled corpus, previously unknown sentiment terms are extracted and added to the initial lexicon.

The remainder of this paper is structured as follows: Section 2 gives an overview of related work, followed by a description of the proposed method in Section 3. Section 4 performs a comprehensive evaluation of our approach, comparing the semi-automatically created lexicons to lexicons assembled by language experts. Section 5 concludes the paper and outlines future work.

2. RELATED WORK

This paper introduces an approach to combine games with a purpose and a lexicon-based sentiment detection method to create domain-specific sentiment lexicons. The following two subsections discuss related work in the field of sentiment

detection, and provide background material on the use of crowd-sourcing applications in the tradition of games with a purpose.

2.1 Sentiment Detection

Sentiment detection heavily relies on so-called sentiment lexicons, i.e. collections of terms and an a-priori assessment of their polarity. Well-known English resources are the General Inquirer [19], the Subjectivity Lexicon [29] and the Subjectivity Sense Annotations [27, 8]. GermanPolarityClues [26] or the lexicon presented by Clematide and Klenner [3] are good examples of equivalent German resources.

Sentiment lexicons are valuable resources, and much work focuses on the creation of such lexicons. This task usually involves a lot of handicraft, making it time-consuming and resource-intensive. This explains the strong interest in reliable automatic approaches.

In an early approach, Hatzivassiloglou and McKeown [9] used syntactical relations to identify new sentiment terms. Turney and Littman [23] use Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) to identify sentiment terms in a large Web corpus. Terms with sufficient co-occurrence frequency with one of 14 paradigm terms (i.e., a gold standard list of seven positive and negative terms) are assigned the same sentiment value as the respective paradigm term. Evaluated on the General Inquirer [19], PMI shows results comparable with the algorithm of Hatzivassiloglou and McKeown [9]. Using three different extraction corpora, Turney and Littman show that PMI does not outperform Hatzivassiloglou's and McKeown's algorithm but is more scalable [24]. LSA provided better results, but was not as scalable as PMI. Turney [22] uses the same techniques to identify new sentiment terms from a paradigm list of only two terms (*excellent* and *poor*). This procedure performed well on the review corpus. Beineke et al. re-interpret the previously discussed mutual association as a Naïve Bayes approach [2]; they also expand this unsupervised approach and create a supervised approach using labeled data.

In Esuli et al. [5] a semi-supervised approach creates *SentiWordNet*, a sentiment resource based on the well-known linguistics resource WordNet [6]. They first manually label all synsets containing 14 seed terms, which results in an amount of 47 synsets with positive label and 58 with negative. All synsets obtained from certain relations (e.g. *direct antonymy*, *similarity* and *derived-from*) with these seed synsets are labeled accordingly. Synsets without connection to the seed sets are classified as objective, as long as they do not have a different sentiment value in the General Inquirer. The so gathered data is used to train eight ternary classifiers, which classify the rest of WordNet. Kim and Hove [10] specify subjects by means of a Named Entity Recognition and assign them the overall sentiment value of the sentence. A list of 44 verbs and 34 adjectives expanded by WordNet synonyms and antonyms serves as sentiment lexicon. A straightforward solution to accomplish sentiment detection in a language without existing sentiment lexicon is to use translation software. Denecke [4] applies a machine learning approach to multi-lingual sentiment detection using movie reviews from six different languages. Google Translate (www.google.com/language_tools) converts foreign-language documents into English. The feature selection procedure extracts a total of 77 features out of four super classes [4]: (i) the frequency of word classes (i.e. the number of verbs, nouns, etc.); (ii) polarity scores for the 20 most frequent words and the averages scores for all verbs, nouns and adjectives are based on SentiWordNet [5]; (iii) the frequency of positive and negative words according to the General Inquirer; and (iv) textual features such as the number of question marks.

The a priori polarity of sentiment terms might change in different contexts. This problem is tackled by Gindl et al. [7], proposing an approach that dynamically refines the polarity by invocation of context. The first step is the identification of ambiguous terms in a sentiment lexicon. For each of these ambiguous terms, probabilities for their occurrence in positive and negative contexts are calculated by analyzing their occurrence in a corpus of positive and negative reviews. Based on this information, the a priori polarity of an ambiguous term is modified by analyzing terms co-occurring with the term in an unknown lexicon. Wilson et al. [29] examine 28 syntactical and linguistic features in a machine learning approach. Several of those features are context-based, e.g. invoking the sentence preceding or succeeding the current one or the document topic. The features are tested using BoosTexter's AdaBoost.MH algorithm [15] on the Multi-perspective Question Answering Opinion Corpus [28]. The approach has two steps: the first step filters subjective sentences from objective ones, and the second assigns sentiment values to the subjective sentences. In their successive work [30] Wilson et al. use four different machine learning algorithms to test their feature selection and also use a larger version of the corpus. Agarwal et al. [1] use the corpus to test n-grams and provide syntactical label for relations as context characteristics. Polanyi and Zaenen propose context handling strategies from a linguistic perspective [12]. They distinguish two main groups of context modifiers: *Sentence Based Contextual Valence Shifters* and *Discourse Based Contextual Valence Shifters*.

Please refer to the surveys by Liu [11] and Tang et al. [21] for a more exhaustive overview of sentiment detection.

2.2 Games with a Purpose

Human language technologies such as information extraction and sentiment detection depend on appropriate language resources. Such resources can be acquired through Games with a purpose [25, 14], a crowd-sourcing mechanism and a special type of serious games that invites communities of users with different levels of expertise to participate in value-adding processes. Games with a purpose leverage collective intelligence, which is described as combining "behavior, preferences, or ideas of a group of people to create novel insights" [17]. Collective intelligence from groups of people often produces better results than individual domain experts [20].

Games with a purpose have been used successfully to solve problems that computers cannot yet solve, such as tagging images [25] and annotating content [18]. The main challenges of game design are motivating users to play the game while generating useful data, and ensuring that the process yields unbiased results. Given appropriate design and authentication mechanisms, such games can capture individual knowledge according to the scientific criteria of objectivity, reliability, validity and representativeness. In the context of this paper, we harness the wisdom of the crowds through games with a purpose to be delivered via large-scale social

networking platforms such as Facebook for compiling multilingual sentiment lexicons. Advantages of this approach include a large number of possible players, intrinsic motivation within a social context, and more effective mechanisms to detect and combat attempts of manipulating results. When adopting an approach based on filtering and cross-validation, the intrinsic motivation of users participating in games with a purpose promises superior results compared to crowd-sourcing marketplaces such as Amazon Mechanical Turk (www.mturk.com). Merging several types of games (e.g. sentiment lexicon creation, translation, conflict resolution) further increases the game's attractiveness, reduces the risk of cheating, allocates collective intelligence more efficiently by prioritizing tasks across game types, and helps avoid the situation that dedicated players run out of new challenges.

3. METHOD

Sentiment detection techniques use text features such as sentiment terms and sentiment indicators to assess the polarity (positive, negative) of text fragments. *Sentiment terms* have a distinct polarity and are usually domain-independent. In contrast, *sentiment indicators* occur within the discussion of *topics* which are often used in a positive or negative context (e.g. democracy, public debt, etc.). Therefore, these terms do not contain a polarity by themselves but rather *indicate* that the topic is likely to contain a certain sentiment. This is particularly useful in situations where only rudimentary sentiment lexicons are available (e.g. for less spoken languages or unusual application domains), since sentiment indicators have the potential to considerably improve the accuracy of sentiment detection in such settings (Section 4). Nevertheless, since topics are usually domain-specific, sentiment indicators still have the limitation of being specific to a particular domain and, therefore, cannot be used across domains.

The proposed method introduces an approach which automatically extracts *sentiment terms* and *sentiment indicators* by applying a bootstrapping process to domain-specific documents. The retrieved indicators then complement sentiment dictionaries and increase the sentiment detection's recall.

The sentiment values of *domain-specific sentiment terms* are usually limited to a particular domain. *Sentiment indicators* such as "democracy" or "tax raise" do not contain a sentiment value per se but are associated with a certain sentiment in the given domain. Therefore, they provide a good indication of how an article is going to be perceived by its readers.

One objective of our approach is to improve the recall of sentiment detection for languages where sentiment resources are limited or still under development.

The presented approach starts with the creation of an initial sentiment lexicon as described in Subsection 3.1. Based on this lexicon a bootstrapping algorithm (see Subsection 3.2) extracts further sentiment terms and indicators used to expand the initial lexicon.

3.1 Initial Sentiment Lexicon

This paper builds upon the lessons learnt from the Sentiment Quiz (Figure 2), a Web-based social verification game for sentiment detection. It was developed as part of the US Election 2008 Web Monitor (www.ecoresearch.net/election2008), a project to investigate information diffusion via interactive online media, and the interdependence of news media coverage and public opinion [16].

The game is available in seven different languages and presents the player with potential sentiment terms. The player's task is to evaluate these terms on a five-point scale (very positive, positive, neutral, negative, very negative) and he receives points based on how well his answer corresponds to the other player's assessment of a particular term. If no prior evaluations are available for a term, the game assigns the player a score which is based on his average game performance. The sentiment quiz attracted more than 4 300 players who have created a sentiment lexicon comprising 1 000 high quality terms as a by-product of their activities.

Figure 2: The Sentiment Quiz, a word polarity game (www.modul.ac.at/nmt/sentiment-quiz)

A crucial task when applying such games with a purpose is to make sure that the games yield unbiased results and that users are prevented from raising their score by cheating. On a social networking site, users can identify other players and might collaborate to manipulate the game; e.g. by agreeing in advance on the answers to a limited set of questions. A number of simple measures can be taken to ensure output of high quality: (i) hide the identity of the other player; (ii) analyze the temporal distribution of answers; (iii) assign trust values to each player, which in turn determine the impact of their answers – e.g. insert questions with known answers into the exercise queue and identify users who tend to score low on these questions; (iv) avoid exploitable patterns in the sequence of answers, since users who identify the pattern could quickly earn credits without actually solving the puzzle. We also only consider terms which have received at least seven assessments to ensure a good quality of the initial sentiment lexicon used for the bootstrapping process.

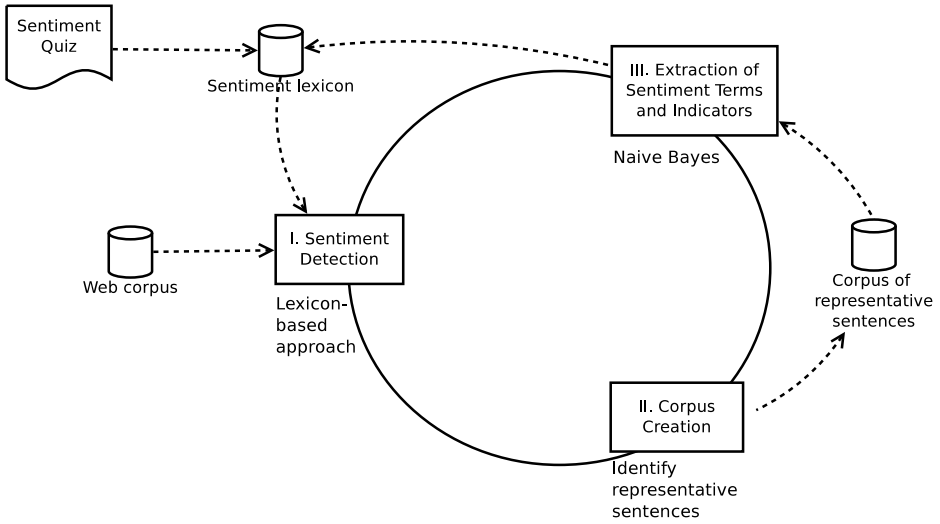

Figure 1: The three-step bootstrapping process

3.2 Bootstrapping Algorithm

We apply a bootstrapping algorithm to extract potential sentiment terms and sentiment indicators for the given domain. An unlabeled corpus of TripAdvisor reviews serves as input for this step.

Figure 1 proves an overview of the three-step bootstrapping process. Initially we apply sentiment detection to determine the sentiment of unlabeled Web reviews (Section 3.2.1) based on an initial sentiment lexicon, which was created by crowd-sourcing the task of annotating vocabulary with sentiment values to a Facebook game with a purpose (Section 3.1). We then identify representative examples of reviews with a positive and negative sentiment and use them to create a corpus of such reviews (Section 3.2.2). Finally, we extract sentiment indicators and terms from this corpus (Section 3.2.3), merge these terms into the sentiment dictionary, and repeat the process as required.

3.2.1 Sentiment Detection

Applying a simple lexicon-based sentiment detection approach estimates the sentiment (σ) of the extracted reviews:

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

$$n(t_{i-j}) = \begin{cases} -1.0 & \text{if } t_{j-1} \text{is a negation trigger} \\ +1.0 & \text{otherwise} \end{cases} \quad (2)$$

The algorithm uses a bag of words approach and considers negation by scanning for negation triggers such as 'not" and "without" which invert the sentiment value of the following term. We applied a simple lexicon-based approach, which only considers simple grammatical constructs such as negation, for detecting the sentiment of unlabeled documents. For the evaluation we complemented this approach with a Naïve Bayes classifier and Support Vector Machines.

3.2.2 Corpus Creation

The next step creates and expands a corpus of positive and negative reviews to be used for the extraction of sentiment terms and indicators. The output of the sentiment detection component helps to identify the k strongest positive and negative reviews (doc_i) and the corresponding sentiment thresholds (σ_k^+ and σ_k^-). Due to the strength of their sentiment values we consider these reviews as *representative examples* of positive and negative discussions and therefore assemble corresponding learning corpora containing positive C^+ and C^- negative examples:

$$C^+ = \{doc_i | \sigma(doc_i) > \sigma_k^+\} \quad (3)$$

$$C^- = \{doc_i | \sigma(doc_i) < \sigma_k^-\} \quad (4)$$

The input corpus is a collection of 1 600 unlabeled holiday reviews downloaded from the website www.tripadvisor.com. The corpus is balanced, containing an equal number of positive and negative reviews. We assign a positive polarity when a review has more than three stars, and a negative if it has less than three stars.

3.2.3 Extraction of Sentiment Terms and Indicators

The extraction of new sentiment terms follows each expansion of the corpora (C^+ and C^-). For each term in the knowledge base the system calculates its probability of occurring in positive and negative sentences based on the Naive Bayes algorithm.

$$n(t_j) = n(t_j|C^+) + n(t_j|C^-) \quad (5)$$

$$P(\sigma(t_j)|C^+) = \frac{n(t_j|C^+)}{n(t_j)} \quad (6)$$

$$P(\sigma(t_j)|C^-) = \frac{n(t_j|C^-)}{n(t_j)} \quad (7)$$

Subsequently, the m terms with the highest absolute probability values and the corresponding sentiment thresholds P^+ and P^-, i.e. the strongest m positive and negative terms, are added to the sentiment lexicon. Terms already included in the lexicon are disregarded. We also ignore terms which occur less then n_{min} times in the corpus.

$$\sigma(t_j) := 1 \quad \text{if } P(\sigma(t_j)|C^+) > P^+ \wedge n(t_j) \geq n_{min} \tag{8}$$

$$\sigma(t_j) := -1 \quad \text{if } P(\sigma(t_j)|C^-) > P^- \wedge n(t_j) \geq n_{min} \tag{9}$$

Our current approach applies this bootstrapping process multiple times and divides the number of representative sentences to include in the corpus creation step (k) by half after every run. The terms yielded by this process include relevant sentiment indicators and sentiment terms which considerably improve the performance of subsequent sentiment detection steps (Section 4).

4. EVALUATION

Figure 3 visualizes the described evaluation process. The evaluation design focuses on the following research questions: (1) is the quality of the bootstrapped and newly included sentiment terms high enough to improve the overall quality of the system, and (2) how well does this lexicon compare to a manually compiled lexicon which was assembled by language experts.

To answer these two questions we performed a 10-fold cross-validation of the following three lexicons based on three different sentiment detection algorithms:

- **The Facebook lexicon:** This lexicon is the result of the Sentiment Quiz described in Section 3.1. It includes 500 positive and 500 negative terms. The game delivered more terms, but we excluded unreliable terms, i.e. we only took those 500 positive and negative terms with the smallest standard deviation from the average assessment of the players
- **The expanded lexicon:** This lexicon is an expansion of the Facebook lexicon. It contains additional terms identified with the bootstrapping algorithm described in Section 3. The system included 127 new terms on average (to accomplish a 10-fold cross validation we had to create an expanded lexicon for each run of the validation to avoid pollution of training data with test data).
- **The General Inquirer lexicon:** This lexicon builds upon the sentiment information contained in the General Inquirer (see Stone [19]). It contains 3 625 sentiment terms in total, 2 006 are negative and 1 619 positive terms.

The corpus used for cross-validation is a collection of 1 600 reviews downloaded from the TripAdvisor website (www.tripadvisor.com). For each run of the cross-validation the system creates an expanded lexicon from the training data. The presented lexicons are used by three different algorithms:

- **Lexical approach:** This algorithm uses a bag of words approach and simple grammar rules (Equation 1 and Equation 2) to determine text sentiment.
- **Naïve Bayes:** The terms in the lexicons serve as features for the Naïve Bayes classifier.
- **Support Vector Machine (SVM):** The lexicon terms also serve as features for the SVM classifier, which uses a linear kernel.

We chose Naïve Bayes and SVM as classifiers since they are standard algorithms and especially SVMs are known to deliver excellent results on high-dimensional data such as textual data. The WEKA tool serves as framework for the evaluation with the Naïve Bayes and the SVM algorithm. For this purpose we first converted the textual reviews into ARFF files, the common file format for WEKA. The lexical algorithm processes the reviews in plain text format. In order to ensure equivalence of the training and test data for both the WEKA environment and the lexical approach we did not use WEKA's built-in 10-fold cross-validation mode but created the corresponding files ourselves.

Tables 1 and 2 contain the results of our evaluation. Table 1 compares the Facebook lexicon with the expanded lexicon. The table can be read as follows: each triple contains the average of either recall, precision, or F-measure achieved with one of the three algorithms using either the Facebook or the expanded lexicon. $\mathbf{R_f}$ refers to the average recall achieved with the Facebook lexicon (f), $\mathbf{R_e}$ refers to recall obtained with the expanded lexicon (e). The column **Sig** has a check mark (✓) when the difference is statistically significant and a dot (·) when it is not. In case the expanded lexicon delivers significantly worse results the column contains a dashed circle (⊖). The R implementation of Wilcoxon's rank sum test serves for calculation of significance values [13]. We regard significance values below 5 % (i.e. $p < 0.05$) as significant.

Table 1: Results of the 10-fold cross-validation with the WEKA LibSVM classifier

Polarity	$\mathbf{R_f}$	$\mathbf{R_e}$	Sig	$\mathbf{P_f}$	$\mathbf{P_e}$	Sig	$\mathbf{F_f}$	$\mathbf{F_e}$	Sig
Lexical									
Positive	77	90	✓	62	69	✓	68	78	✓
Negative	29	43	✓	85	92	✓	43	58	✓
Naïve Bayes									
Positive	63	76	✓	75	79	✓	68	77	✓
Negative	79	79	·	68	76	✓	73	78	✓
SVM									
Positive	73	80	✓	75	79	✓	74	79	✓
Negative	75	78	·	74	80	✓	74	79	✓

Table 2: Comparison of the expanded lexicons with the General Inquirer

Polarity	$\mathbf{R_e}$	$\mathbf{R_{gi}}$	Sig	$\mathbf{P_e}$	$\mathbf{P_{gi}}$	Sig	$\mathbf{F_e}$	$\mathbf{F_{gi}}$	Sig
Lexical									
Positive	90	95	·	69	65	✓	78	77	·
Negative	43	36	✓	92	93	·	58	52	·
Naïve Bayes									
Positive	76	85	⊖	79	82	⊖	77	83	⊖
Negative	79	81	·	76	85	⊖	78	82	⊖
SVM									
Positive	80	86	·	79	82	·	79	84	·
Negative	78	81	·	80	85	·	79	83	·

Table 2 contains a comparison of results achieved with both the expanded lexicon and the General Inquirer lexicon. The results show, that although the semi-automatically

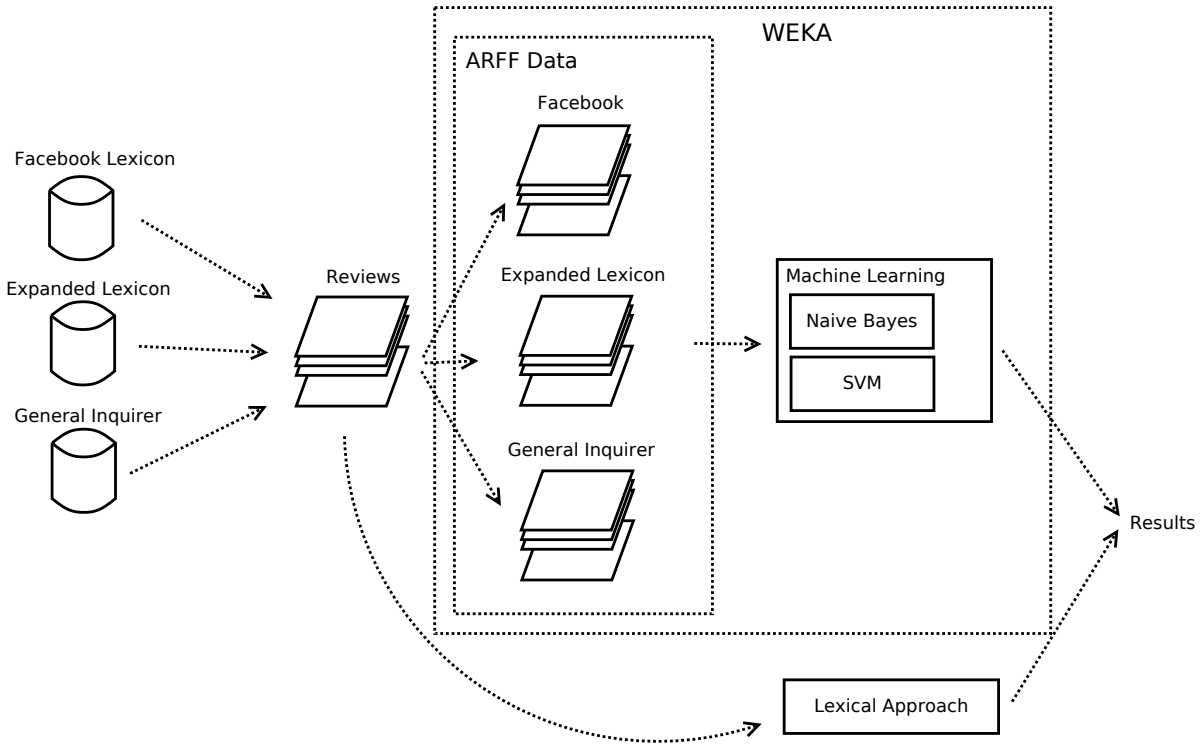

Figure 3: Overview of the evaluation process

compiled sentiment lexicon has less than half the number of sentiment terms, it sill performs similarly to the expert lexicon for two of the three evaluated sentiment detection approaches. The General Inquirer lexicon is only significantly better for results achieved with the Naïve Bayes classifier. We did not observe significant differences for the SVM classifiers, yet the different values still indicate better results of the General Inquirer lexicon. For the lexical approach the expanded lexicon was even able to significantly outperform the General Inquirer lexicon in two cases (precision for positive reviews and recall for negative reviews).

The lexical approach profited the most from the bootstrapping process. We obtained significant improvements for recall, precision, and F-measure. The improvements achieved with the Naïve Bayes and SVM classifiers were all significant except for recall of negative reviews.

Table 3 shows three terms which were incorporated into the sentiment lexicon during the bootstrapping process and lists sentences that illustrate how these terms improve the method's accuracy. Interestingly, the intuitively negative term *stops* was identified as a positive sentiment term. After the lookup of sentences in the databases that contained this sentence, the reason became apparent. The term *stops* referred to bus or subway stations. In general, it is desirable to live close to a bus stop, and the system also identified it correctly. Therefore, *stops* can be considered as one of the afore-mentioned sentiment indicators. Only in the domain of holiday reviews it gets an obvious positive connotation (although it might also be used positively in domains completely different to holiday reviews). The two other examples, *dingy* and *stained* are sentiment terms - one can easily imagine them to be used negatively in a different domain. The significant improvement achieved with the bootstrapped lexicon shows that the proposed method is a valuable tool under circumstances where sentiment resources are sparse.

5. CONCLUSION

This paper proposed a semi-automated process which combines Games with a purpose and a bootstrapping approach to create sentiment lexicons and customize them to a particular domain. Complementing crowd-sourcing with bootstrapping yields an extended sentiment lexicon (containing sentiment terms *and* sentiment indicators), which considerably outperforms the accuracy of the initial dictionary.

The main contributions of this paper are (i) the introduction of the concept of sentiment indicators, which supports sentiment detection by complementing known sentiment terms with domain knowledge, (ii) applying Games with a Purpose to the task of generating language resources which are essential for many natural language detection and knowledge management tasks, (iii) introducing a bootstrapping process which automatically extends these resources by adding sentiment indicators and sentiment terms based on unlabeled domain documents, and (iv) performing a comprehensive evaluation which shows that bootstrapping considerably improves the performance of the created sentiment lexicon, and that the lexicon yielded from the semi-automatic process performs - depending on the used sentiment detection method - about as good or only slightly worse than widely used language resources such as the General Inquirer, which have been compiled by language experts.

Table 3: Examples of terms added after bootstrapping

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

This result is remarkable for a semi-automatically created resource, especially when considering that the main benefit of the introduced method is its applicability to languages and domains for which such high quality resources are not yet available. In such cases the effort required to create language resources is reduced significantly.

The evaluation also demonstrates that the introduced bootstrapping process is very efficient in learning sentiment terms and indicators. Nevertheless, it currently has the disadvantage of not being able to distinguish between domain-independent sentiment terms and topic-related sentiment indicators. This is not a problem for domain-specific sentiment detection as such, but is highly relevant for the ability of reusing sentiment lexicons across domain. Future research will address this shortcoming by applying corpus-based methods such as the one introduced in Gindl et al. [7] for identifying domain-specific sentiment indicators.

We will also explore the applicability of Games with a Purpose to the creation of other language resources such as test collections and text annotations.

6. REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. McKeown. Contextual Phrase-level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams. In *12th Conference of the European Chapter of the Association for Computational Linguistics on (EACL 2009)*, pages 24–32, Athens, Greece, 2009. ACL.

[2] Philip Beineke, Trevor Hastie, and Christopher Manning. Exploring Sentiment Summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 12–15, 2004.

[3] Simon Clematide and Manfred Klenner. Evaluation and Extension of a Polarity Lexicon for German. In *1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010)*, pages 7–13, Lisbon, Portugal, 2010.

[4] Kerstin Denecke. How to Assess Customer Opinions Beyond Language Barriers? In *3rd International Conference on Digital Information Management (ICDIM 2008)*, pages 430–435, London, UK, November 2008. IEEE.

[5] Andrea Esuli, Fabrizio Sebastiani, and Via Giuseppe Moruzzi. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, Genoa, Italy, 2006.

[6] Christiane Fellbaum. WordNet - An Electronic Lexical Database. *Computational Linguistics*, 25(2):292–296, 1998.

[7] Stefan Gindl, Albert Weichselbraun, and Arno Scharl. Cross-Domain Contextualisation of Sentiment Lexicons. In *19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 771–776, Lisbon, Portugal, August 2010. IOS Press.

[8] Yaw Gyamfi, Janyce Wiebe, Rada Mihalcea, and Cem Akkaya. Integrating Knowledge for Subjectivity Sense Labeling. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics on (NAACL 2009)*, pages 10–18, Boulder, CO, USA, 2009. ACL.

[9] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the Semantic Orientation of Adjectives. In *8th Conference on the European Chapter of the Association for Computational Linguistics (EACL 1997)*, pages 174–181, Madrid, Spain, 1997. ACL.

[10] Soo-Min Kim and Eduard Hovy. Determining the Sentiment of Opinions. In *20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 2004. ACL.

[11] Bing Liu. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

[12] Livia Polanyi and Annie Zaenen. *Computing Attitude and Affect in Text: Theory and Applications*, chapter Contextual. Springer, Netherlands, 2006.

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[14] Walter Rafelsberger and Arno Scharl. Games with a Purpose for Social Networking Platforms. In *20th ACM conference on Hypertext and Hypermedia (HT 2009)*, pages 193–198, New York, NY, USA, 2009. ACM.

[15] Robert E. Schapire and Yoram Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39(2):135–168, 2000.

[16] Arno Scharl and Albert Weichselbraun. An Automated Approach to Investigating the Online Media Coverage of US Presidential Elections. *Journal of Information Technology & Politics*, 5(1):121–132, 2008.

[17] Toby Segaran. *Collective Intelligence - Building Smart Web 2.0 Applications*. O'Reilly, 2007.

[18] K. Siorpaes and M. Hepp. Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems*, 23:50–60, 2008.

[19] Philip J. Stone. *The General Inquirer: A Computer Approach to Content Analysis*. M.I.T. Press, Cambridge, Massachusetts, U.S.A., 1966.

[20] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Little, Brown, London, 2004.

[21] Huifeng Tang, Songbo Tan, and Xueqi Cheng. A Survey on Sentiment Detection of Reviews. *Expert Systems with Applications*, 36(7):10760–10773, 2009.

[22] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *40th Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 417–424, Philadelphia, PA, USA, 2002.

[23] Peter D. Turney and Michael L. Littman. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. Technical report, National Research Council, Canada, Institute for Information Technology, 2002.

[24] Peter D. Turney and Michael L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346, 2003.

[25] L. Von Ahn. Games with a Purpose. *Computer*, 39(6):92–94, 2006.

[26] Ulli Waltinger. GERMANPOLARITYCLUES: A Lexical Resource for German Sentiment Analysis. In *7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1638–1642, Valletta, Malta, 2010.

[27] Janyce Wiebe and Rada Mihalcea. Word Sense and Subjectivity. In *Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL 2006)*, pages 1065–1072, Sydney, Australia, 2006. ACL.

[28] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210, 2006.

[29] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, pages 347–354, Vancouver, B.C., Canada, 2005. ACL.

[30] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, 35(3):399–433, 2009.