UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

TOPIC-RELATED SENTIMENT ANALYSIS FOR DISCOVERING CONTRADICTING OPINIONS IN WEBLOGS

Kerstin Denecke, Mikalai Tsytsarau, Themis Palpanas
and Marko Brosowski

July 2009

Technical Report # DISI-09-037

Topic-related Sentiment Analysis for Discovering Contradicting Opinions in Weblogs

Kerstin Denecke
L3S Research Center
Hannover, Germany
denecke@L3S.de

Mikalai Tsytsarau
University of Trento
Trento, Italy
tsytsarau@disi.unitn.eu

Themis Palpanas
University of Trento
Trento, Italy
themis@disi.unitn.eu

Marko Brosowski
L3S Research Center
Hannover, Germany
brosowski@L3S.de

ABSTRACT

This work addresses the problem of analyzing the evolution of community opinions across time. First, a two-step approach is introduced to determine a continuous sentiment value for each topic discussed in a text based on SentiWordNet as lexical resource. Sentences are clustered according to their topic using Latent Dirichlet Allocation. Both steps are extensively evaluated and tested. The output is then exploited for studying contradictions among weblog posts and comments. We introduce a novel measure for contradictions based on a mean value and the variance of opinions among different posts. In addition, a method is proposed, which identifies posts with contradicting opinions on certain topics on a basis of such a measure. It can be used to analyze and track opinion evolution over time and to identify interesting trends and patterns. The developed algorithm is applied to a dataset of medical blogs and comments on political news with promising performance and accuracy.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles; H.3.1 [**Information Systems**]: Information Storage and Retrieval—*Content Analysis and Indexing*

General Terms

Blog Analysis, Sentiment Analysis, Topic Detection

1. INTRODUCTION

In the last few years, blogs have become more and more popular as they are offering a new medium for users to present information, express opinions and get feedback from other users. Weblogs collectively represent a rich source of information on a person's life in general, but more importantly on a myriad of different topics, ranging from politics and health to product reviews.

It is now becoming evident that the views expressed in blogs can be influential to readers in forming their opinions on some topic [15]. Similarly, the opinions recorded in blogs are an important factor taken into consideration by product vendors [14] and policy makers [24].

In this paper, the focus is on discovering topics automatically for which different opinions have been expressed across space and time. In this way, an interactive analysis and an online identification of contradictions under multiple levels of time granularity is possible. Evolution of opinions can be tracked over time and interesting trends and patterns can be identified. Consider the following motivating scenarios.

Health: The medical blogging community is composed of both trained and certified physicians and individuals from the general population (e.g., patients, or relatives of patients). The blogs written by physicians allow patients and doctors to form an idea about physician's opinions on current health topics, and also how these views evolve over time. In contrast, the personal perspective of patient blogs allows physicians to learn about the mental, emotional and physical state of people living with certain medical conditions and how these change over time.

Politics: Political blogs cover the entire spectrum of interested parties: from simple citizens expressing their opinions on everyday issues, to politicians using this medium in order to communicate their ideas (as was best exemplified during the last US elections), and from journalists criticizing the government to the government itself. It is to the benefit of all the parties mentioned above to follow the opinions that are expressed on a variety of topics in these blogs, and to be able to identify how these opinions or public sentiment change and evolve across time.

In both scenarios, one of the most intriguing aspect worth investigating further is when the opinion of an individual or a group of people on a specific topic changes from positive to negative, or vice-versa. These contradictions may signify a change of mind in the way a certain disease is treated, or may indicate a change of direction of the government with respect to some political issue.

The techniques we describe in this paper are focused on the tasks of extracting opinions from weblogs organized by topic, and efficiently identifying and analyzing contradicting opinions. We make the following contributions.

- Introduction and evaluation of a sentiment analysis ap-

proach that identifies the sentiment per topic in a more granular manner.

- Application and evaluation of topic models to cluster sentences according to their topic.
- Introduction of an approach to detect contradictions in opinions regarding specific topics.
- Experimental evaluation on two real-world datasets, related to health and politics.

The rest of this paper is organized as follows. Section 2 discusses related work. We formally define the problems in Section 3, and describe our solutions in Sections 4 and 5. The results of the experimental evaluation are discussed in Section 6. We conclude and discuss directions for future work in Section 7.

2. RELATED WORK

In the following paragraphs, we briefly discuss the related work in the areas of topic identification, sentiment extraction, and contradiction analysis.

2.1 Topic Identification

In this paper, blog contradictions are considered at topic-level, i.e., topics per blog post need to be discovered. To solve this task, different topic representation and detection methods are available, such as clustering of documents based on extracted keywords [31] or filtering of documents using networks of relations between tags [27]. The TopCat system [7] exploits natural language processing techniques to identify key entities in texts and then forms clusters with a hypergraph partitioning scheme. Substitution of topic identification with a lexicon look-up to determine product names, person names and the like as topics within the opinion mining task has been proven successful for processing specifically product or movie reviews [16, 29]. In addition, most of the existing research determines topics at document-level. Since we analyze sentiments on sentence-level, our approach also has to determine topics at this level.

The most relevant approach to our work is the work of Mei et al. [22], who propose a probabilistic topic sentiment model. Our approach differs in that sentiment calculation and topic detection are performed successively. Topic models are used to identify topics at sentence-level and the sentiment is represented on a continuous scale.

2.2 Opinion Mining and Sentiment Extraction

In existing research work, sentiment analysis is mostly considered as two- or three-class classification problem, distinguishing between *positive* or *negative* (or *neutral*) texts. Different lexical- and machine-learning approaches have been developed [26], e.g., using corpus statistics [32] or linguistic tools like WordNet [17]. The algorithms were mainly applied to movie [1] or product reviews [8].

Our approach goes beyond the classical classification problem and tries to assign a continuous value to a sentence reflecting the expressed opinion. The sentiment analysis task considered in this paper is most similar to the rating inference task in which the class labels are scalar ratings such as 1 to 5 "stars" representing the polarity of an opinion. Rating inference tasks were by now considered at document level [25] or on product feature-level [19, 28]. Pang and Lee [25] apply metric labeling to assign a value of a rating scale while Shimada and Endo [28] use frequency of words as classification features. Ku et al. [18] determine polarity scores between -1 and 1 indicating the polarity and the strength of a word. For this purpose, they calculate the frequency of single characters in positive and negative words of their opinion dictionary.
In contrast to existing rating inference approaches, our algorithm assigns a continuous value to each sentence or topic. Therefore, this task cannot be considered as multi-class classification problem. SentiWordNet [11] provides for each synset of WordNet[1] a triple of polarity scores (positivity, negativity and objectivity) whose values sum up to 1. It has been created automatically by means of a combination of linguistic and statistic classifiers and consists of around 207000 word-sense pairs or 117660 synsets. Existing work exploits this resource mainly for identification of opinionated words [10, 12]. In contrast, our approach will rely on SentiWordNet scores themselves as calculation attributes.

2.3 Contradiction Analysis

A traditional approach in obtaining trends for popular items in blogosphere is to track user support for a set of popular keywords, i.e., measuring the frequency of keywords. Glance et al. describe BlogPulse [13], a system for identifying trends in weblog entries. This method uses frequency as a measure of popularity and relevance, but does not focus on how opinions may vary. Chi et al. [5] introduce a Singular Value Decomposition method for the analysis of trends in topic popularity across time. Some research work also examines how sentiments in blog entries of a single user change over time [21]. The problem of identifying and analyzing opinions has also been studied in the context of social networks. A recent study [6] examines how communities in blogosphere transit between high- and low-entropy states across time, incorporating sentiment extraction. Varlamis et al. [30] propose clustering accuracy as an indicator of blogosphere opinion convergence.

Closer to our work is the analysis of opinions expressed about commercial products, which has attracted particular attention in the research community. Morinaga et al. [23] describe a system for mining the reputation of products in the web. A similar approach is proposed by the Opinion Observer system [20] that focuses on summarizing the strengths and weaknesses of a particular product. Even though the above studies consider both positive and negative opinions, they do not aggregate them. In our approach, we describe an effective way for performing this aggregation, which leads to more insights into user opinions.

Chen et al. study precisely the problem of conflicting opinions [4] on a corpus of book reviews, which they classify as positive and negative. Their main goal is to identify the most predictive terms for the above classification task, and visualize the results for manual inspection. In contrast, we propose a systematic and automated way of performing opinion aggregation, revealing contradictions, and analyzing the evolution of these contradictions over time.

3. PROBLEM DEFINITION

The problem we want to solve in this paper is to detect contradicting opinions on certain topics and to analyze their evolution across time in the blogosphere. Evidently, in order

[1]http://wordnet. princeton.edu/

to identify contradictions, we first have to solve the problems of topic extraction and sentiment analysis. In the rest of this section, we elaborate on the above issues, and formally define the problems we address in this study.

3.1 Definition of Relevant Terms

Usually, a particular blog covers some general topic (e.g., health, politics) and has a tendency to publish more posts about one topic than another. Yet, within a blog post, the author may discuss several different specific topics.

DEFNINITION 1 (BLOG POST TOPIC). *A topic T is a named entity, event or abstract concept that is described in a blog post, P. We refer to all the topics contained in a single post as P* topics*, $\mathcal{T}^P$. Similarly, the blog posts that refer to a specific topic T are the T* posts*, $\mathcal{P}^T$.*

For each of the topics discussed in a blog post, we wish to identify the author's opinion or sentiment towards it. In this study, we restrict ourselves to identifying and recording the *polarity* of these sentiments, which we represent as numbers. In addition to computing the sentiment polarity on a particular topic given an individual reference to it, we also need to compute the polarity on that topic aggregated over multiple posts (that may span different authors, as well as time periods). In the following, we refer to sentiment polarity simply as *sentiment*, and to the polarity of sentiments aggregated over a collection of posts as *topic sentiment*[2].

DEFNINITION 2 (SENTIMENT). *The sentiment S on topic T in a post P is a real number in the range $[-2, 2]$ that expresses the author's opinion on T. Negative values indicate negative opinions and positive values represent positive opinions.*

DEFNINITION 3 (TOPIC SENTIMENT). *The Topic Sentiment S^T of a collection of posts $\mathcal{P}^T$, which are published within some predefined time window w on topic T, is defined as the aggregated value of the sentiments expressed in $\mathcal{P}^T$ with respect to T.*

In this work, we use the range of $[-2, 2]$ to represent sentiment values, though, in principle any other range could be used as well. As will become evident later on, expressing sentiments using a continuous range of values gives us flexibility in aggregating and analyzing them.

We now turn our attention to the issue of comparing the sentiment values of different collections of posts.

DEFNINITION 4 (SIMULTANEOUS CONTRADICTION). *In a collection $\mathcal{P}^T$ of posts talking about topic T, the topic T is considered contradictory, if there exist two groups of posts $\mathcal{P}_1^T, \mathcal{P}_2^T \subset \mathcal{P}^T$ such that the sentiment S_1 of $\mathcal{P}_1^T$ is very different to the sentiment S_2 of $\mathcal{P}_2^T$.*

In the above definition, we purposely not specify exactly what it means for a sentiment value to be very different from another one. This definition can lead to different implementations, and each one of those will have a slightly different interpretation of the notion of contradiction. We believe that our definition captures the essence of contradiction, without trying to impose any of the particular interpretations. Though, later on (in section 5) we propose a specific method for computing contradictions, which incorporates many desirable properties.

[2]For the rest of this document we will use the terms *sentiment* and *opinion* interchangeably.

Another interesting situation arises when the majority of posts within some time interval exhibits a positive (negative) sentiment on a particular topic, and this time interval is followed by another one, where the majority of posts exhibits a negative (positive) sentiment on the same topic. Such time intervals, that contain a change of topic sentiment, can also be identified as contradictory, but with a special type of contradiction, which we call *Change of Sentiment*.

DEFNINITION 5 (CHANGE OF SENTIMENT). *We say that we have a change of sentiment for topic T, at time t, when the following condition is satisfied: $\exists$ time interval $\tau : \forall \epsilon \leq \tau : S^T(t-\epsilon)S^T(t+\epsilon) \leq 0$.*

3.2 Formulation of Problems

In order to detect contradicting opinions in collections of posts, we first need to determine all the different topics that appear in the posts, and calculate the sentiment of these topics. Subsequently, we can detect the contradicting topics that appear in the dataset.

PROBLEM 1 (TOPIC IDENTIFICATION). *Identify a set $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ of topics of interest that are discussed in the set $\mathcal{P} = \{P_1, P_2, \ldots, P_i\}$ of blog posts.*

PROBLEM 2 (SENTIMENT EXTRACTION). *For a topic $T \in \mathcal{T}^P$ in blog post P, we identify the sentiment S that has been expressed by the author on T in P.*

PROBLEM 3 (CONTRADICTING TOPIC DETECTION). *For a given time interval τ, and topic T, identify the time windows w, contained in τ, where a simultaneous contradiction or a change of sentiment occurs for T, with contradiction values above some threshold ρ.*

The time interval, τ, is user-defined. As we will discuss later, the threshold, ρ, can either be user-defined, or automatically determined in an adaptive fashion based on the data under consideration.

The approach we propose in this work is general, and can lead to solutions for several variations of the above problem, such as detecting the topics with the highest contradiction or the most frequently contradicting topics. For the sake of brevity, in this paper we will demonstrate only the first solution.

4. SENTIMENT EXTRACTION

The algorithm for analyzing contradictions works in two steps: First, for each topic discussed in a blog post, a sentiment value is calculated. Then, the actual contradiction analysis takes place. The methods for topic detection and sentiment analysis work on sentence level. Their results are later aggregated to come up with topic-sentiment pairs for the most relevant topics within one post. The different methods are described in the following sections.

4.1 Identification of Topics

For identifying topics per sentence, we apply the Latent Dirichlet Allocation algorithm (LDA, [2]), which was initially implemented to cluster complete documents according to their topic. We extended the algorithm by a sentence detection algorithm to apply it to sentences that are then considered as input 'documents' for the LDA.

First, each post is splitted into sentences using the sentence splitting library provided by Lingpipe[3] and a regular

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

expression. Lingpipe is a framework for linguistic analysis of human language that offers java libraries for text classification and linguistic processing. The regular expression corrects bad formatted endings of sentences including sentences without proper space, dots, and question marks.

In a second step, topics along with their probabilities are identified for each sentence using the LDA-algorithm and based on the vector representation of sentences. Each sentence is considered to consist of a topic mixture and each word's creation is attributable to one of the sentence's topics. Therefore, a topic is described by a set of words derived from the documents where to each word a probability is assigned that indicates the relevance of this word for the topic. In this way, all topics are described by the same words, but with varying probability values for each word. The number of topics has to be fixed at the beginning of the clustering process.

In our evaluation, we consider the top 3 words to be most relevant for describing a topic. Nevertheless, it can occur, that none of the top 3 topic words can be found within a sentence to which this topic has been assigned. The LDA algorithm is not looking for matching keywords, but it is creating a model that describes the topic. The benefit of these topic models is that the correct topic can be assigned even if no matching keyword occurs in the sentence, just by relying upon a larger set of words, or the context, respectively. We ran the LDA with standard parameters for α and β as reported by Steyvers and others who found that $\alpha = \frac{50}{t}$ and $\beta = 0.01$ where t is the number of topics work well with many different text collections [3].

In order to exclude sentences without topical focus, our LDA modification considers only sentences with at least four words (excluding stop words). A term is in turn considered relevant for clustering when it occurs in at least 15 sentences. By now, words are neither normalized morphologically nor stemmed. We are currently studying the influence of such a preprocessing to the quality of topic detection. The probability per topic and sentence calculated by LDA indicates to what degree the sentence belongs to the topic. In our approach, only topics with a probability larger than 0.75 are considered for further processing. Topics with a smaller probability are excluded since their support for this topic is too low. One reason for choosing LDA are these probability values that allow us to filter out irrelevant topics. In future, we will study the influence of this threshold to the quality of topic detection.

4.2 Identification of Opinions

For each (relevant) topic determined by the modified LDA algorithm (see above), a continuous value between -2 and 2 is assigned indicating the sentiment expressed regarding this topic in the sentence under consideration. SentiWordNet [11] already provides continuous values representing the polarity of single words. Thus, we decided to develop an approach based on this resource. Other existing approaches (as discussed in Section 2.2) assign a numeric value or distinguish only between *positive* and *negative* texts and are therefore not directly applicable to our scenario.

4.2.1 Sentiment Calculation

The polarity scores provided by SentiWordNet can be used in two different ways. We propose a rule-based approach, but also a machine-learning based approach. Both approaches require the calculation of average polarity triples for words of each sentence of a post.

Rule-based sentiment calculation: The sentiment regarding a topic is determined based on relevant opinionated words. Words are considered relevant if they appear close to a topic term, i.e. within a distance of four words before and after the topic term. We are exploiting the top 3 topic words for this purpose. In case none of the top 3 topic words can be found in the sentence under consideration, all words of a sentence are considered for sentiment detection. Stop words are removed, the resulting words are stemmed and their polarity score triples are collected from SentiWordNet. These values are in turn averaged which results in one polarity score triple per sentence. By calculating the difference between positivity and negativity value of this triple, the final continuous topic-related sentiment value is determined. Since SentiWordNet scores are in range between 0 and 1, we use a scaling factor of 2 to receive values in our sentiment range. If the resulting value is smaller than -2 (or larger than 2), the polarity value is set to -2 (or 2). Objectivity values are not considered in this rule since we only want to account for opinionated topics. In this way, to each topic of a sentence a sentiment value between -2 and 2 is assigned.

Machine-learning based sentiment calculation: Another possibility to calculate a polarity value for each sentence is by applying machine learning techniques, in particular logistic regression models. In this case the assumption is used that a sentence's polarity corresponds to the topic's polarity. For applying regression models, classification features are necessary.

We exploit a feature set that has previously been used to classify complete texts as *positive* or *negative* [9]. It consists of the number of positive, negative and neutral words, the number of adjectives, verbs and nouns, as well as the SentiWordNet triples of the five most frequent terms. For each sentence, words are tagged with their part of speech using the QTagger[4]. The single words of a sentence are stemmed and looked up in SentiWordNet. For each synset matching the stemmed word, negativity values or positivity values are summed up and the average value is calculated. This results in a polarity score triple for each term of a sentence. The SentiWordNet score triples are exploited to count the *positive* and *negative* words within a sentence. If the positivity value of a term is larger than the negativity value, the word is considered to be *positive* and *negative* otherwise. If both values are equal, the word is considered to be *neutral*. In addition, the SentiWordNet score triples of the five most frequent terms provide an additional extension to the feature set.

The resulting 21 feature values are used by Linear Regression models. Regression is the process of computing an expression that predicts a numeric quantity. In our evaluation, the WEKA implementation of regression models LinearRegression is used [33]. Linear regression starts with numeric attributes to predict a numeric outcome. The idea is to express the class as a linear combination of the attributes with predetermined weights. The weights are calculated from the training data. Using this algorithm, a sentiment value is assigned to each sentence.

[4]Available at: http://www.english.bham.ac.uk/staff/omason/software/qtag-api.html

4.2.2 Assigning Sentiments to Document Topics

The previously described steps provide for each sentence of a post sentence-topic-sentiment triples. The topic-sentiment values per sentence are aggregated to determine one sentiment value for each topic of a post. For this purpose, the sentiment values of sentences with the same topic are averaged. The final output of the sentiment analysis step is a continuous polarity value between -2 and 2 for each topic of a post.

The main contribution of the sentiment analysis approach is determining the semantic orientation of a sentence in a more fine-grained manner using SentiWordNet. We decided to determine topics and sentiments at sentence-level to be able to consider changes of sentiment within one post. It may occur that regarding one topic different opinions are expressed in different sections of the same post. So, we have to identify all the word expressing the opinion towards this topic. For example, there is a WebMD post, where the author states as a fact that there are discussions on over-prescription of a certain drug. The matching topic key word occurs only in this sentence. In the other sentences, he collects arguments in favor and against this statement, but resists on repeating the relevant topic keywords. By considering sentiment per sentence and relating it to the topic, as it is proposed by our approach, we are able to detect these different opinions regarding the same topic and to aggregate them.

5. CONTRADICTION ANALYSIS

Based on the analysis described so far, we are now in position to detect the contradicting topics. In the following paragraphs, we first propose a novel contradiction measure, and then describe a simple, yet effective way of organizing the data in order to identify contradictions based on this measure.

5.1 Measuring Contradiction

In order to be able to identify contradicting opinions we need to define a measure of contradiction. Following Definition 3, the topic sentiment for topic T can be calculated as the mean value of the opinions of all the posts that mention T, $\mathcal{P}^T$: $S^T = \frac{1}{n}\sum_{i=1}^{n} S_i$, where n is the cardinality of $\mathcal{P}^T$. Then, a value of S^T close to zero implies a high level of contradiction.

A problem with the above way of calculating topic sentiment arises when there exists a large number of posts with very low sentiment values (i.e., values close to zero). In this case, the value of S^T will be drawn close to zero, without necessarily reflecting the true situation of the contradiction. Therefore, we suggest to additionally consider the variance of the sentiments along with their mean value.

Defnınition 6 (Topic Sentiment Variance). *In a collection $\mathcal{P}^T$ of posts talking about topic T, the topic sentiment variance V_S^T is defined as follows: $V_S^T = \frac{1}{n}\sum_{i=1}^{n}(S_i - S^T)^2$.*

According to the above definition, when there is a large uncertainty about the collective sentiment of a collection of posts on a particular topic, the topic sentiment variance is large as well.

Figure 1 shows two example sentiment distributions. Distribution A with S^T close to zero and a high variance indicates a very contradictive topic. Distribution B shows a far less contradictive topic with sentiment S^T in the positive range and low variance.

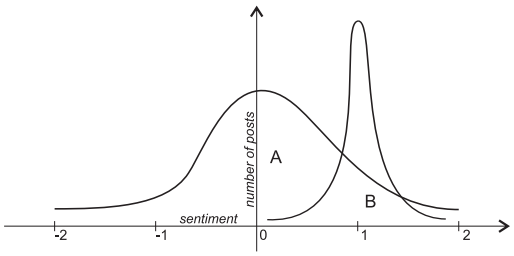

Figure 1: Example of two possible sentiment distributions.

Evidently, we need to combine topic sentiment and topic sentiment variance in a single formula for computing contradictions. Assume that we want to look for contradictions in a shifting time window w. Without loss of generality, in this work we consider windows of a day, week, month, and year. For a particular topic T, the set of posts $\mathcal{P}^T$ will be restricted to those, that were posted within window w. We denote this set of T posts as $\mathcal{P}^T(w)$. Then, the contradiction value C^T can be computed as $C^T = \frac{V_S^T}{(S^T)^2}$, where S^T is squared so that its units are the same as the units of V_S^T.

This formula captures the intuition that contradiction values should be higher for topics whose sentiment value is close to zero, and sentiment variance is large. Nevertheless, the contradiction values generated by this formula are unbounded (i.e., they can grow arbitrarily high as S^T approaches zero), and does not account for the number of posts in $\mathcal{P}^T(w)$. This latter point is important, because in the extreme where $\mathcal{P}^T(w)$ contains only two posts with opposite values, C^T will be very high, and will compare unfavorably to the contradiction value of a different set of T posts with a much higher cardinality.

Incorporating to the contradiction formula the observations made above, we propose the following final formula for computing contradiction values: $C^T = \frac{V_S^T}{\alpha+(S^T)^2}W$. In the denominator, we add a small value, $\alpha \neq 0$, which allows to limit the level of contradiction when $(S^T)^2$ is close to zero. In this study, we used a value of $\alpha = 0.05$, which was effective for its purpose, without distorting the final results.

W is a weight function aiming to compensate the contradiction value for the varying number of posts that may be involved in the calculation of C^T. The weight function is defined as $W = 2 + \tanh\left(\frac{n}{10} - 3\right)$, where n is the cardinality of $\mathcal{P}^T(w)$. This weight function is a multiplicative factor in the range $[1, 3]$ (Figure 2 plots W as a function of n), which means that contradiction values fall within the interval $[0, 12/\alpha]$. Using W we can effectively limit C^T when there is a minor number of posts, as well as when this same number of posts increases significantly. What W achieves is essentially a normalization of the contradiction values across different sets of T posts, allowing them to be meaningfully compared to each other.

Figure 3 demonstrates the operation of the proposed contradiction value function. The graph at the top (Figure 3(a)) shows a time series of synthetically generated sentiments for a period of 8000 time units. The dataset consists of 4000 normally distributed opinions with dispersion 0.5 and median following a custom trend. Additionally, we added 4000 points of normally distributed sentiments with dispersion 1 and median 0, acting like noise. Time stamps of all points

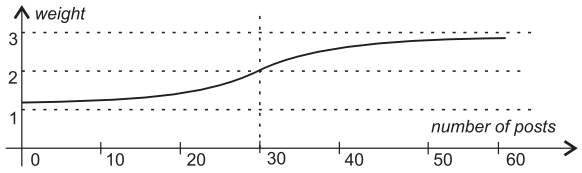

Figure 2: The weight function used with number of posts in the criteria.

follow the Poisson distribution with parameter $\lambda = 2$ time units. The bold line in this graph depicts the custom trend, showing an initial positive sentiment that later changes to negative (at time instance t_1). This behavior represents a change of sentiment. There is also a point around time instance t_2, where the sentiments are divided between positive and negative, a situation representing a simultaneous contradiction. Using this dataset, we verify the ability of the C^T function to capture the planted contradictions.

An important component of C^T is the topic sentiment, S^T. As can be seen in Figure 3(b), S^T closely captures the aggregate trend of the raw sentiments. The following two graphs in the figure show the contradiction value, calculated using a sliding window of size 500 and 1000 time units. When we use a window of small size (Figure 3(c)), C^T correctly identifies the two contradictions at points t_1 and t_2, where the values of C^T are the largest. Using a larger window has a smoothing effect in the values of C^T (Figure 3(d)). Nevertheless, we can still identify long-lasting contradictions: In this case, the largest value of C^T occurs at time instance t_1, corresponding to a change of sentiment that manifests itself across the entire dataset. The above observations also indicate the value of examining contradictions using time windows of varying cardinality. In the following paragraph, we describe how this can be done efficiently.

5.2 Identification of Contradictions

Our final goal is to identify contradictions in large collections of posts. To this end, we organize the sentiment information on each topic across different time windows that form a time hierarchy, namely, days, weeks, months, and years. We subsequently compute the contradiction values for each topic and time window. This organization allows us to identify contradictions on any topic, in any of the above time windows. The information is stored in a relational database, following the schema shown in Table 1 (in the full version of this paper, we discuss efficient techniques for the incremental update and maintenance of this table, by storing some key statistical sums that still allow the exact computation of the C^T values). Contradiction values for each topic with respect to time intervals of different granularities for all topics are stored within the same table, leading to simple and efficient SQL queries for detecting the interesting contradictions. These queries can return all topics with a contradiction value greater than some threshold in a particular time interval τ, as well as change of sentiment[5].

In the above discussion, we assume that the user is interested in all contradictions above some fixed threshold[6].

[5]Note that when the user specifies a time interval, the algorithm returns contradictions for all windows in the time-hierarchy contained in that interval.

[6]Extensions to identifying top-k contradictions are straightforward, and outside the scope of this work.

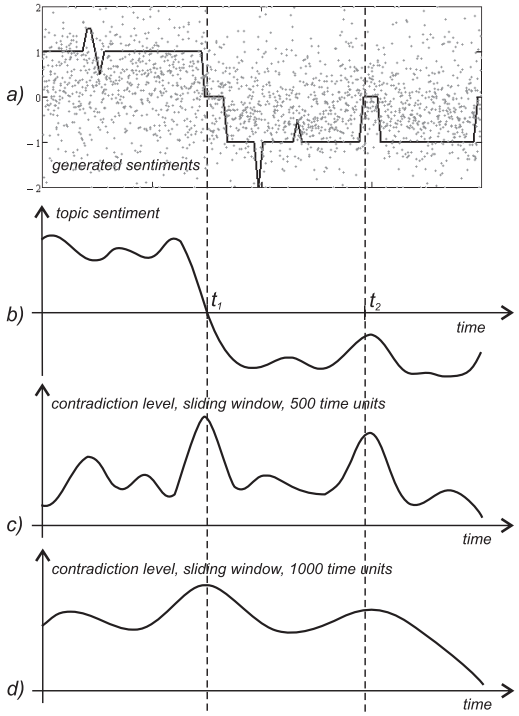

Figure 3: Example of contradiction values computed from a synthetic dataset with two planted contradictions.

Attribute Name	Description
topicId	Topic identifier
timeBegin	Timestamp of the interval beginning
timeEnd	Timestamp of the interval end
granularity	A level of granularity
contradiction	Pre-computed contradiction value

Table 1: A schema for the table containing summary values.

Alternatively, we can set the threshold level independently for each time window based on the value of contradiction of the parent window (in the time hierarchy). A simple solution is to set the threshold to p times the contradiction value of the parent (for the same topic), where $0 < p < 1$. In this study, we use $p = 0.9$. This method allows the contradiction threshold to better adapt to the data. If there is a small value of contradiction at a higher granularity, the threshold becomes smaller. We can thus detect interesting contradictions that occur in different time granularities and across topics, even if these contradictions do not have the largest values overall. This is particularly important when a single, fixed threshold value cannot detect all contradictions across time, or when the user is unsure about which threshold to choose.

For the sake of brevity, we describe the SQL expressions of the relevant queries in the full version of this paper.

6. EXPERIMENTAL EVALUATION

In this section, we evaluate the steps of contradiction analysis. Starting with an introduction of the corpus, we evaluate topic and sentiment extraction, as well as the performance and accuracy of the contradiction detection algorithm.

Annotators	A vs. B	B vs. C	C vs. D
Agreement	46.6%	44.8%	73.5%
Mean Error	0.797	0.852	0.394

Table 2: Agreement of annotators.

6.1 Corpus Description

Our algorithms are applied to a data set of health-related weblog posts from WebMD and a dataset with comments on postings from Slashdot (http://slashdot.org).

We crawled 28 health-related blogs with 2,405 posts covering 4 years (January 2005 to January 2009) from the WebMD webpage (http://blogs.webmd.com/). These posts are written by health care professionals and report on certain health topics such as disorders (e.g., *sleep disorders, asthma, anxiety*) or certain treatments (e.g., *cancer treatments*, *cosmetic surgery*).

Slashdot is a popular website for people interested in reading and discussing about technology and its ramifications, and includes posts, as well as comments on these posts. In this study, we used a dataset provided for the CAW2 workshop (http://caw2.barcelonamedia.org/) that contains about 140,000 comments under 496 articles, covering the time period from August 2005 to September 2006.

For the sentiment analysis problems considered in this paper, there are no publicly available annotated text corpora. Therefore, we created such a corpus by randomly extracting 500 sentences from the Slashdot dataset. These sentences were annotated at sentence-level by four humans with a continuous value between -2 and 2 representing the polarity of the sentences. We hired two undergraduate students of computer science without any background knowledge in sentiment analysis (labeled annotator A and B), and two postdoctoral researchers working in the fields of natural language processing and semantic web services (labeled annotator C and D). None of them is an author of this paper. We don't expect an exact agreement in the assigned sentiment values. Instead, values of two annotators for one sentence are considered as an agreement, when their difference is smaller than 0.5.

The results are given in Table 2. It can be seen that the agreement between the annotators A and B, as well as B and C, is restricted to around 45%. In contrast, annotator C and D agree for 73.5% of the sentences. The table also shows the mean absolute error which indicates, to what degree the assigned values differ. The smallest mean error was 0.394 for agreement of annotators C and D. The other annotator pairs disagreed to a larger extent. Obviously, the sentences under consideration can be interpreted in different ways resulting in largely differing sentiment values in annotations. The assigned opinion values of the different annotators vary a lot, indicating the difficulty of this task. To take this observation into account, we created the gold standard for the evaluations in section 6.3 by calculating the average of the sentiment annotations per sentence. Annotators A and B agree with these average values at about 54%, while C and D agree on about 66% and 70% of the cases.

In addition, a dataset comprising reviews on 14 products from amazon.com (e.g., mobile phones, mp3 player, digital camera, Diaper) provided by Minqing Hu and Bing Liu (http://www.cs.uic.edu/~liub) is used for evaluation purposes. These reviews have been annotated on sentence level with product features and associated opinion. A positive opinion is represented by a numeric value of 1, 2 or 3, while a negative opinion is represented by -1, -2 or -3. Since time stamps are missing in that particular dataset, the contradiction approach could not be applied. But, we use this data set in our evaluations of topic and sentiment analysis.

Identified Topics (Slashdot)		Identified Topics (WebMD)	
Rank	Keywords	Rank	Keywords
example topics for LDA-20		**example topics for LDA-20**	
1	china chinese people	1	bad cancer pain
2	country american people	2	clinical drug research
3	free market people	3	child health care
4	internet government control	4	medical surgery doctor
5	companies company money	5	heart disease blood
6	article read gov	6	fat food grams
7	bush court president	7	based air advice
8	day election war	8	tea cup green
9	people car fact	9	time day sleep
10	law laws case	10	body brain cell
example topics for LDA-200		**example topics for LDA-200**	
1	companies market internet	1	knee replacement pain
2	congress law constitution	2	eye surgery vision
3	bush administration clinton	3	fat grams calories
4	argument free copyright	4	healthy diet fat
5	anti god evolution	5	blood heart disease
6	access cd music	6	breast cancer age
7	attack iraq war	7	depression stress
8	government amendment	8	birth control credit
9	china economy people	9	ear infections infection
10	people civil democracy	10	coffee caffeine food

Table 3: Some of the identified topics for Slashdot and WebMD datasets.

6.2 Evaluation of Topic Detection

6.2.1 Evaluation Methodology

For evaluating the quality of the LDA on sentence-level, two people were asked to evaluate the topic keywords determined by the LDA algorithm for 500 sentences of the customer review data set and of the Slashdot data set. The test persons were confronted with the three most probable topic terms of the three topics with the highest probability. They had to select the topic that describes best the topic of the sentence under consideration based on these words. LDA was run with 20 and 200 topics.

Although the customer reviews are already annotated with product features on sentence level, these annotations are not well suited for our evaluation since LDA proposes more general terms as topic terms (e.g., *camera, router*) than the assigned product features.

6.2.2 Evaluation Results

For both data sets similar evaluation results are achieved. When LDA clustered the sentences into 20 topics, the annotators marked topic words of only 21% of the sentences as correct. For the other sentences none of the suggested topic terms were relevant. Significantly better results were achieved when we increased the number of topics used by LDA. In our tests, when the number of topics was set to 200, 54% of the topics assigned to sentences were marked as correct by the annotators. Table 3 reports examples of topics identified for both the Slashdot and the WebMD datasets, when using LDA with a limit of 20 and 200 topics. Apart from the relevance of results, we observe that LDA with 200 topics is also able to identify topics that are more specific.

The LDA algorithm fails when misleading words were in a sentence. For example the sentence *taking pics of my 7-month old baby* was obviously assigned to the topic *diaper* because the sentence is talking about a baby. Since the sentences are often very short, the information is sometimes

correct	Text
Topic "back pain arthritis"	
yes	I don't mind seeing patients with back pain.
yes	I have my daily aches and pains related to arthritis (a family legacy).
no	The last time this happened, I pushed to get back to work.
Topic "government law federal"	
yes	This is a right only as long as it's backed up by the power of the government.
yes	Unfortunately, the Commonwealth of Virginia has taken the exact opposite tact.
no	Sounds crazy that they'd agree to sign a contract like that.

Table 4: Example sentences with detected topics.

insufficient for clustering correctly (e.g. the comment for a camera *sound is not loud enough* leads to the topic (terms) *ipod* and *songs*).

The evaluation shows that preprocessing of the sentences need to be improved: We expect quality improvements when performing stemming or a morphological analysis within a preprocessing step of the LDA and consider synonyms. This would help to cluster sentences with the same words in plural or singular similarly (e.g. *cards* vs. *card*) and also to consider synonyms appropriately (*Microsoft* vs. *MS*). In addition, we are currently testing the quality of LDA when restricting topic words to nouns.

In Table 4 some example sentences with assigned topics are listed. It can be seen that some sentences are correctly assigned, some even if none of the top 3 topic keywords are contained (e.g., second sentence for the topic 'government law federal'). On the other hand, for some sentences assigned topics seem to be unrelated to the topic (e.g., negative example for topic 'back pain arthritis'). The latter shows that the algorithm fails, when words are used in a different context (e.g., *back* in *get back* instead of *back pain*).

6.3 Evaluation of Sentiment Analysis

6.3.1 Evaluation Methodology

The accuracy of the proposed approach for sentiment analysis is determined based on the 500 sentences that have been manually annotated (see section 6.1). The average of the values assigned to each sentence by the four annotators are used as ground truth. We calculated the 'error' which is the difference between two annotations for the same sentence. The 'mean absolute error' is calculated by averaging the 'errors' of all annotated sentences. In addition, we determine the accuracy which we define by the percentage of agreements of two annotators, where two annotated values are considered as an agreement if the 'error' for the sentence under consideration is smaller than 0.5. We compared the mean absolute error for the proposed rule-based and machine-learning based approaches. Furthermore, modifications of the approaches are tested. T-Tests are made to ensure statistical significance of the results.

6.3.2 Rule-based Approach

Table 5 shows the results when applying the approach to the evaluation material. The best mean error rate of 0.366 is achieved when resisting on stemming, i.e., words remain unstemmed when looking for SentiWordNet scores. This modification achieves with a confidence of 85% better results than the original method. Compared to the annotator agreements with the ground truth presented above, our approach achieves a slightly larger agreement than the annotators C and D with the ground truth.

Modification	Accuracy	Mean absolute error
stemmed	70.9%	0.368
unstemmed	72%	0.366

Table 5: Results for the rule-based approach.

Sentiment	Text
Examples with correct assignments	
1.6	Practitioners and patients alike swear by the effectiveness of particular healing methods, even where there may not be a scientific explanation of how they work or even empirical evidence that they do really work.
1.5	It's easy to find really good examples of sensible taxation in the US.
0.3	Face it, there are things in the world that justify a society taking action to regulate it for the better of the society.
-2.0	Both of these nasty arachnids can cause a painful bite, tissue damage, and even death.
-1.5	Something went wrong during the anesthesia and her little, normal brain was irreversibly damaged.
-1.2	That there are ways to violate this law with little risk or that the law is bad and/or unenforceable is utterly irrelevant to his point, which is that the penalties prescribed are way out of line with the seriousness of the "crime".
Examples with wrong assignments	
0.6	I believe that information technology is important, but I think that MIT is just trying to get publicity, something the Media Lab specializes in (Added nasty putdown - the Media Lab doesn't do very good science or engineering in my opinion.).
0	Is Cambridge indirectly helping the Chinese government to fix firewall issues?
1.5	Valentine's Day has become an event filled with pressure to love on demand – and that's the very antithesis of romance or good sex.

Table 6: Example sentences with sentiments detected by the rule-based approach.

It is interesting to note that values assigned by our approach differ from the ground truth most when also the human annotators disagreed to a large extent. For example, for the sentence *What about those inside China using those exploits for legitimate ends?* the manual annotations are 0.4, 0.3, -0.7 and -1 (average: -0.25). The automatically assigned value is slightly positive with 0.28.

Another problem is due to the stemming of words that can lead to wrong polarity triples collected from SentiWordNet. For example, the synsets for the terms *ironic* and *iron* are considered after stemming as belonging to the same term *iron*. In this way, the negative meaning of the term *ironic* becomes rather objective since there are words with neutral meanings (*iron, ironing*). The results showed that resisting on stemming helps to improve the results. A future extension of the approach would be to select the correct synset based on corpus statistics.

Often, the misclassified sentences require background knowledge to correctly decide for a sentiment value which is missing in our current approach. E.g., for interpreting the sentence *Many Canadians themselves leave the country in what the government refers to as a 'brain drain'.* correctly, background knowledge is necessary. The approach also fails, when rather neutral words are used to express a *positive* or *negative* opinion e.g., *Isn't Cambridge deliberately creating an opportunity for the Chinese government to prosecute them?*. The last sentence in the examples in Table 6 shows that the algorithm also fails when negativity is expressed very subtly.

6.3.3 Machine learning-based approach

In addition to the rule-based approach, we tested the machine-learning based approach, that assigns a continuous value based on SentiWordNet scores. We tested two different feature sets, namely, the feature set as described in Section 4.2, and a second feature set that only consists of

the average SentiWordNet scores of the three word classes adjectives, nouns and verbs.

The Linear Regression classifier was trained on the customer reviews described in Section 6.1. The assigned values of this data set were multiplied in advance by $\frac{2}{3}$ to have values in the same range as used by our algorithm and of the ground truth. The test set was again provided by the 500 annotated examples from the Slashdot data set (see Section 6.1). The two feature sets achieve similar accuracies of around 70% and mean absolute errors of 0.39. Compared to the rule-based approach, the machine-learning approach performs with similar accuracy. But, a clear benefit of using a rule is that no training material is necessary, which in our case is rather sparse.

Since sentiment analysis is often domain dependent and in our evaluation training and test set were of different text types and domains, we also studied the results of the machine-learning based approach for the customer review data set only. 90% of the data set was reserved for training the classifier; 10% were used for testing. An accuracy of 60.4% and a mean absolute error of 0.61 could be achieved. These results are significantly lower than those achieved for the Slashdot sentences. In a larger evaluation, we will investigate the domain-dependency of this approach.

6.4 Evaluation of Contradiction Analysis

Finally, we apply the introduced contradiction analysis approach to the WebMD and the Slashdot dataset.

In Figure 4, the top graph depicts the raw sentiment values for the topic "internet government control" (from the Slashdot dataset), for the time interval January to September 2006. The following graphs show the topic sentiment and variance (two middle graphs), and contradiction values (bottom graph) for the above topic and time interval. Contradiction values have been calculated using a time window of one day. Note that contradiction values are high for the time windows where topic sentiment is around zero and variance is high, which translates to a set of posts with highly diverse sentiments. These situations are not easy to identify with a quick visual inspection of the raw sentiments.

The analysis shows that in this time interval there are three major contradictions (marked 1-3 in the bottom graph of Figure 4). All three contradictions discuss the pros and cons of a law that would give the government more power in controlling the internet traffic, especially personal correspondence. By taking a closer look at the corresponding weblog posts, we find out that the discussion around contradiction 1 is about web-related corporations operating in a monopolic or oligopolic environment, and how the above law would affect their operation and their customers. Contradiction 2 contains discussion of how this law can or cannot deter attacks from foreign countries, while contradiction 3 discusses the ways that this law will or will not affect national and foreign citizens.

Evidently, these are all very relevant discussions that express different points of view on the same topic, and having an automated way of identifying them can be very useful. Table 7 shows extracts from two opposing posts that contributed to contradiction 2. In the same table, we also report additional examples of contradictions identified by our analysis for two more topics.

Finally, we evaluated the time-performance of our database approach for detecting contradictions. Figure 5 illustrates

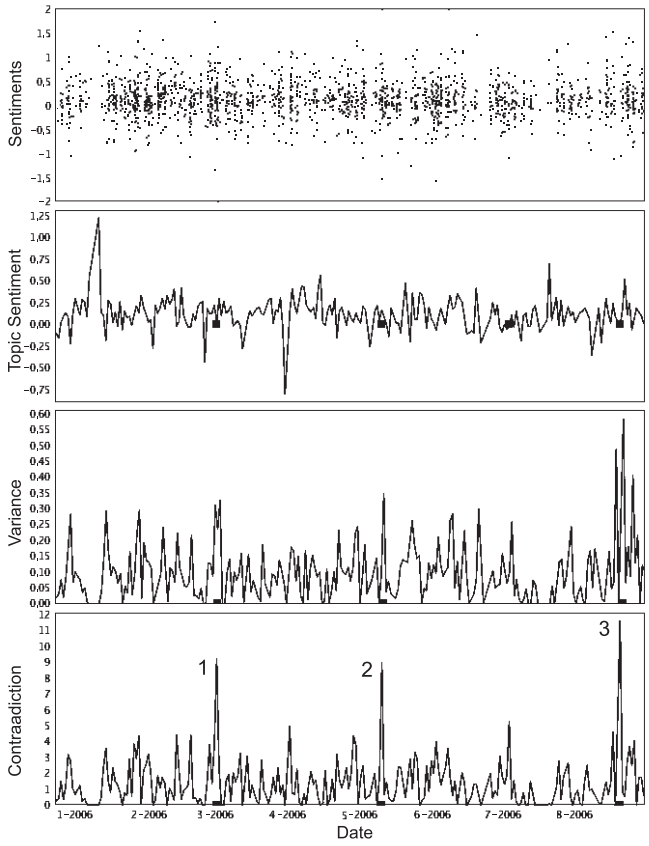

Figure 4: Raw sentiment, topic sentiment, topic sentiment variance, and contradiction values for the topic "internet government control".

the time needed to query the database for all contradicting topics (above a threshold), as a function of the time interval, τ, within which the contradiction has to occur. Remember that the database reports relevant contradictions for all windows in the time hierarchy that are contained in the user-specified interval τ. The graph shows that our solution exhibits very good scalability characteristics (almost constant performance) as we increase the interval τ, exploiting the underlying relational database technology. The adaptive threshold queries require more time since the threshold in this case has to be computed by the contradiction value of the parent time window.

Similar results were observed for other types of contradiction analysis queries, even as the number of topics increase. (Note that the number of weblog posts does not affect the performance of contradiction analysis.) Due to lack of space, we defer the detailed discussion of our fully-fledged time-performance evaluation to the full version of this paper.

7. CONCLUSION

In this paper, we study the problem of identifying contradictions in weblog posts. We formally define two variations of this problem, and we introduce a method to detect and analyze such contradictions. The experimental evaluation, with two diverse real-world datasets, demonstrates the applicability of the proposed solution, and shows very promising results with respect to the efficiency and effectiveness of

For the topic "internet government control" (contradiction 2)	
PRO	How about to make a positive impact on the world by gathering and protecting information to prevent terrorists from carrying out acts of violence and to stop hostile countries from threatening the security of the United States and its allies. Because that is what the NSA does!
CON	How do you want to block a top level domain? At the end, you'll find out that those sites will be accessed via the IP address. You're making inappropriate assumptions here.

For the topic "diet fat day"	
PRO	... Any decrease in breast cancer in the experimental group would be measurable by comparing the two groups...
CON	... A low fat diet does not decrease the incidence of invasive breast cancer in post-menopausal women...

For the topic "china america people"	
PRO	Tell you what, China has its own means to deal with issues/problems. It's not prefect, but it's always the most practicable approach. That's China/Chinese surviving skill. You guys can NOT stand on Western foot to judge China, no way, never worked, will never work. China has at least 2500 years history with everything documented. She doesn't need to be told what is right or wrong. America has 300 years history is just equivalent to one dynasty in China. It's just too short to tell who is right for American.
CON	You'd rather live in the Soviet utopia of engineering advances and everyone living in gulags. Or there's the great Chinese vision of an internet without the word freedom on it. The people willing to risk their lives to get off the island of Cuba aren't just fleeing a poor economy, you know? The word intelligence must mean something else in the land you're from.

Table 7: Extracts from posts found in some of the contradicting points.

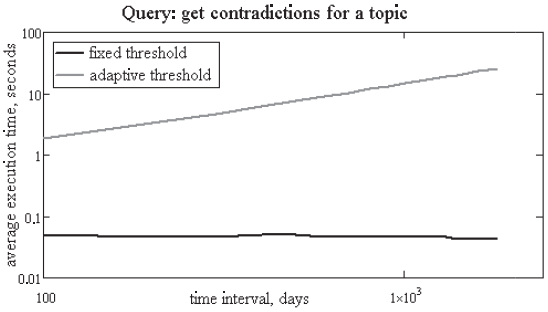

Figure 5: Evaluation of the query execution time for fixed and adaptive threshold approaches.

our techniques.

We are currently working on a more detailed performance evaluation of the contradiction identification algorithm, including scalability tests, which will be reported in a separate paper. In addition, we will study the performance of LDA when considering synonyms in the topic identification algorithm, and when words are preprocessed by a stemming algorithm. Another extension would be to refine the results of the topic identification by domain-specific knowledge (e.g., using an ontology, or a relevant term dictionary), in order to only consider topics that are related to the domain under consideration.

References

[1] M. Annett and G. Kondrak. A comparison of sentiment analysis techniques: Polarizing movie blogs. In *Canadian Conference on AI*, pages 25–35, 2008.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *JMLR*, 3, 2003.

[3] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–48, 2005.

[4] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver. Visual analysis of conflicting opinions. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 59–66, 2006.

[5] Y. Chi, B. L. Tseng, and J. Tatemura. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In *CIKM*, pages 68–77, 2006.

[6] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Multi-scale characterization of social network dynamics in the blogosphere. In *CIKM*, pages 1515–1516, 2008.

[7] C. Clifton, R. Cooley, and J. Rennie. Topcat: Data mining for topic identification in a text corpus. In *IEEE Transactions on knowledge and Data Engineering 16(8)*, pages 949–964, 2004.

[8] K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW*, 2003.

[9] K. Denecke. How to assess customer opinions beyond language barriers? In *Third International Conference on Digital Information Management, ICDIM*, pages 430–435, 2008.

[10] A. Devitt and K. 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, 2007.

[11] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proc of LREC 2006 - 5th Conf on Language Resources and Evaluation*, 2006.

[12] A. Fahrni and M. Klenner. Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In *Proc.of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention, 1st-2nd April 2008. University of Aberdeen, Aberdeen, Scotland*, pages 60 – 63, 2008.

[13] N. Glance, M. Hurst, and T. Tomokiyo. Automated trend discovery for weblogs. In *WWW*, 2004.

[14] T. Hoffman. Online Reputation Management is Hot - But is it Ethical? Computerworld, feb 2008.

[15] J. A. Horrigan. Online shopping. Pew Internet and American Life Project Report, 2008.

[16] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI*, pages 755–760, 2004.

[17] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using wordnet to measure semantic orientations of adjectives. In *Proc LREC 2004*, 2004.

[18] L.-W. Ku, Y.-S. Lo, and H.-H. Chen. Using polarity scores of words for sentence-level opinion extraction. *Proceedings of NTCIR-6 Workshop Meeting*, pages 316–322, 2007.

[19] C. W.-K. Leung, S. C.-F. Chan, and F.-L. Chung. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In *ECAI 2006 Workshop on Recommender Systems*, pages 62–66, 2006.

[20] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW*, pages 342–351, New York, NY, USA, 2005. ACM.

[21] R. McArthur. Uncovering deep user context from blogs. In *AND*, pages 47–54, 2008.

[22] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, pages 171–180, New York, NY, USA, 2007. ACM.

[23] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *KDD*, pages 341–349, New York, NY, USA, 2002. ACM.

[24] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[25] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL-05*, 2005.

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

[27] L. Qu, C. Müller, and I. Gurevych. Using tag semantic network for keyphrase extraction in blogs. In *CIKM*, pages 1381–1382, 2008.

[28] K. Shimada and T. Endo. Seeing several stars: A rating inference task for a document containing several evaluation criteria. In *PAKDD 2008, LNAI 5012*, pages 1006–1014, 2008.

[29] V. Stoyanov and C. Cardie. Annotating topics of opinions. In *International Conference on Language Resources and Evaluation*, 2008.

[30] I. Varlamis, V. Vassalos, and A. Palaios. Monitoring the evolution of interests in the blogosphere. In *ICDE Workshops*, pages 513–518, 2008.

[31] C. Wartena and R. Brussee. Topic detection by clustering keywords. In *DEXA '08: 19th International Conference on Database and Expert Systems Application*, pages 54–58, 2008.

[32] J. Wiebe. Learning subjective adjectives from corpora. In *Proc 17th National Conf on AI and 12th Conf on Innovative Applications of AI*, 2000.

[33] I. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.