Data Mining and Knowledge Discovery manuscript No.
(will be inserted by the editor)

Survey on Mining Subjective Data on the Web

Mikalai Tsytsarau · **Themis Palpanas**

DAMI Special Issue on 10 Years of Mining the Web
Preprint version, accepted for publication on September 08, 2011

Abstract In the past years we have witnessed Sentiment Analysis and Opinion Mining becoming increasingly popular topics in Information Retrieval and Web data analysis. With the rapid growth of the user-generated content represented in blogs, wikis and Web forums, such an analysis became a useful tool for mining the Web, since it allowed us to capture sentiments and opinions at a large scale.

Opinion retrieval has established itself as an important part of search engines. Ratings, opinion trends and representative opinions enrich the search experience of users when combined with traditional document retrieval, by revealing more insights about a subject. Opinion aggregation over product reviews can be very useful for product marketing and positioning, exposing the customers' attitude towards a product and its features along different dimensions, such as time, geographical location, and experience. Tracking how opinions or discussions evolve over time can help us identify interesting trends and patterns and better understand the ways that information is propagated in the Internet.

In this study, we review the development of Sentiment Analysis and Opinion Mining during the last years, and also discuss the evolution of a relatively new research direction, namely, Contradiction Analysis. We give an overview of the proposed methods and recent advances in these areas, and we try to layout the future research directions in the field.

Keywords Sentiment Analysis · Opinion Mining · Contradiction Analysis

M. Tsytsarau
University of Trento, Trento, TN 38100, Italy
Tel.: +39-0461-28-3908
E-mail: tsytsarau@disi.unitn.eu

T. Palpanas
University of Trento, Trento, TN 38100, Italy
Tel.: +39-0461-28-3908
Fax: +39-0461-28-2093
E-mail: themis@disi.unitn.eu

1 Introduction

Since the World Wide Web first appeared two decades ago, it has changed the way we manage and interact with information. It has now become possible to gather the information of our preference from multiple specialized sources and read it straight from our computer screen. But even more importantly, it has changed the way we share information. The audience (i.e., the receivers of the information) does not only consume the available content, but in turn, actively annotates this content, and generates new pieces of information. In this way, the entire community becomes a writer, in addition to being a reader. Today people not only comment on the existing information, bookmark pages, and provide ratings, but they also share their ideas, news and knowledge with the community at large.

There exist many mediums, where people can express themselves on the web. Blogs, wikis, forums and social networks are examples of such mediums, where users can post information, give opinions and get feedback from other users. In their own right, they collectively represent a rich source of information on different aspects of life, but more importantly so on a myriad of different topics, ranging from politics and health to product reviews and traveling. The increasing popularity of personal publishing services of different kinds suggests that opinionative information will become an important aspect of the textual data on the web.

Due to the ever-growing size of the information on the web, we are now barely able to access the information without the help of search engines. This problem gets harder, when we want to aggregate the information from different sources. Multiple solutions have been proposed to solve this problem, and they are mainly specialized in factual information retrieval. To achieve this, subjectivity filtering is applied (Riloff et al, 2005), in order to remove texts that may provide a biased point of view. These texts can be distinguished by analyzing sentiments expressed by the authors, or by discovering explicit marks of contradiction with other texts (Ennals et al, 2010b). This dimension of web search emphasizes the importance of the problem of analyzing subjective data.

We now turn our attention to the following interesting question: whether the subjective data that exist on the web carry useful information. Information can be thought of as data that reduce our uncertainty about some subject. According to this view, the diversity and pluralism of information on different topics can have a rather negative role. It is well understood, that true knowledge is being described by facts, rather than subjective opinions. However, this diversity in opinions, when analyzed, may deliver new information and contribute to the overall knowledge of a subject matter. This is especially true when the object of our study is the attitude of people. In this case, opinionative data can be useful in order to uncover the distribution of sentiments across time, or different groups of people.

It is now becoming evident that the views expressed on the web can be influential to readers in forming their opinions on some topic (Horrigan, 2008). Similarly, the opinions expressed by users are an important factor taken into consideration by product vendors (Hoffman, 2008) and policy makers (Mullen and Malouf, 2006). There exists evidence that this process has significant economic effects (Antweiler and Frank, 2004; Archak et al, 2007; Chevalier and Mayzlin, 2006). Moreover, the

opinions aggregated at a large scale may reflect political preferences (Tumasjan et al, 2010) and even improve stock market prediction (Bollen et al, 2010). These arguments are illustrated in the following examples.

Example 1. Today we can see a growing number of blogs focused on various aspects of politics. They 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 USA presidential 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, and to be able to identify how these opinions or public sentiments change and evolve across time.

Example 2. Imagine a potential buyer of a digital camera, who is not familiar with the details of this technology. In this case, reading the camera specifications can be an arduous task. In contrast, the opinion of the community that shares the same interests with the buyer, can be very informative. Therefore, a system that accumulates feedback and opinions originating from multiple sources, effectively aggregates this information, and presents the result to the user, can be both helpful and influential.

In this study, we introduce readers to the problems of *Opinion Mining* and *Opinion Aggregation*, which have been rapidly developing over the last decade, as well as with a rather new trend related to these areas, namely, *Contradiction Analysis*. In the rest of this document, we will use the term *Subjectivity Analysis* to refer to all three of the above problems together.

The rest of this document is organized as follows. In Section 2 we provide a general view of subjectivity analysis and outline major problems of this domain. Development, problems, definitions and main trends of this area are described in Sections 3 through 5. We analyze and discuss the state of the art in Section 6. Finally, we conclude in Section 7.

1.1 Relation to Previous Work

The interested reader can also refer to previous surveys in the area, among which we point out Pang and Lee (2008) and Tang et al (2009), that helped in the systematic study of opinion and review mining. Our current study has notable differences to the ones mentioned above, with respect to both new content, and also to the way that some common references are being treated.

The Tang et al (2009) survey has a different focus than our study, closely examining the particular problem of sentiment extraction from reviews. When compared to the Pang and Lee (2008) survey, we provide a considerable amount of new information specific to opinion mining: discussions of 26 additional papers, and more extensive discussions for another 7 papers. Moreover, instead of focusing only on the machine learning aspect of the relevant methods and algorithms, we build up our discussion of the opinion mining field around a classification of the papers into four different approaches, following the trends in this field. These approaches are: *machine learning*, *dictionary-based*, *statistical*, and *semantic*. We provide discussions on what the form of the problem that all papers in each one of these four approaches solve is, and, where applicable, we also include a mathematical formulation.

In addition, our study includes new discussions on some topics that have emerged only recently: *opinion mining in microblogs and streaming data*, where we describe 6 studies focused on mining continuously arriving short-length texts; *opinion quality and spam*, where we present 5 techniques that try to detect artificially-constructed opinionated data; and *contradiction analysis*, where we report on 17 approaches, whose focus is on identifying and analyzing conflicting opinions.

Furthermore, this study presents novel comparative information (in the form of graphs and tables) on the algorithms in this area, related to the techniques they use, their performance, as well as to the datasets used for their experimental evaluation. This information helps the reader form an overall picture for the general area of Subjectivity Analysis: where the past efforts have concentrated on, which the most popular methods and techniques are, and what the current trends are.

Finally, we note that this study should be considered as complementary to the earlier surveys in the area, which contain much additional information. The purpose of our study is to highlight the development of the field with a focus on the recent years, examine the main trends that have appeared in the study of the field and their evolution over time, report in a systematic way the performance results of competing algorithms and the characteristics of the available datasets, and discuss some of the emergent topics and open research directions in the area of Subjectivity Analysis.

2 Subjectivity Analysis: A General View

Subjectivity Analysis involves various methods and techniques that originate from Information Retrieval (IR), Artificial Intelligence and Natural Language Processing (NLP). This confluence of different approaches is explained by the nature of the data being processed (free-form texts) and application requirements (scalability, online operation). Therefore, Subjectivity Analysis shares much of its terminology and problem definitions with the domains mentioned above.

The Subjectivity Analysis domain is still in the process of being shaped, and its problem statements touch upon different domains. Being originally studied in different communities, the problems of *Opinion Mining* and *Sentiment Analysis* have slightly different notions. Opinion Mining originates from the IR community, and aims at extracting and further processing users' opinions about products, movies, or other entities. Sentiment Analysis, on the other hand, was initially formulated as the NLP task of retrieval of sentiments expressed in texts. Nevertheless, these two problems are similar in their essence, and fall under the scope of Subjectivity Analysis. For the rest of this document, we will use both these terms interchangeably.

At a first level of approximation, the various Subjectivity Analysis techniques can be described as being composed of the following three steps:

1. *identify*; 2. *classify*; 3. *aggregate*.

These steps also implicitly list the most important problems in Subjectivity Analysis. For example, a typical opinion mining process involves the first two steps, and results in producing sentiment values for texts. In opinion aggregation, the third step is involved as well, in order to aggregate these sentiments. Note that even though this

aggregation can be considered as a post-processing step, it is no less important than the previous steps. Indeed, the analyst is often times more interested in determining the common features and interesting patterns that emerge through sentiments from many different data sources, rather than in the opinions of particular authors.

In the following paragraphs, we discuss in more detail the literature on the problems of *Opinion Mining* and *Opinion Aggregation*. We review the recent developments in these areas, and then present the field of *Contradiction Analysis*, which has recently started to attract interest.

3 Opinion Mining

Opinion Mining is the problem of identifying the expressed opinion on a particular subject and evaluating the polarity of this opinion (e.g., whether the expressed opinion is positive or negative). Opinion Mining forms the basis upon which other tasks under Subjectivity Analysis can be built. It provides an in-depth view of the emotions expressed in text, and enables the further processing of the data, in order to aggregate the opinions, or identify contradicting opinions. Evidently, the quality of the results of Opinion Mining is crucial for the success of all subsequent tasks, making it an important and challenging problem.

3.1 Definitions of Opinion Mining

Opinion Mining operates at the level of documents, that is, pieces of text of varying sizes and formats, e.g., web pages, blog posts, comments, or product reviews.

Definition 1 (Document) Document D is a piece of text in natural language.

We assume that each document discusses at least one topic, and not all topics discussed in the same document have to be related to each other.

Definition 2 (Topic) Topic T is a named entity, event, or abstract concept that is mentioned in a document D.

Usually, a particular information source covers some general topic (e.g., health, politics, etc.) and tends to publish more material about this general topic than others. Yet, within a general topic, the author may discuss several more specific topics[1]. Being able to identify the specific topics is vital for the successful analysis of sentiments, because sentiments are usually attached to them and become their traits.

Examples of such topics are product features, famous persons, news events, happenings, or any other concept s that may attract our interest. What we are interested in is analyzing these topics in connection to any subjective claims that accompany them. Therefore, for each of the topics discussed in a document, we wish to identify the author's opinion towards it.

Definition 3 (Sentiment) Sentiment S is the author's attitude, opinion, or emotion expressed on topic T.

[1] From here on, we will refer to specific topics simply as "topics".

Sentiments are expressed in natural language, but as we will see below, they can in some cases be translated to a numerical or other scale, which facilitates further processing and analysis.

There are a number of differences in meaning between emotions, sentiments and opinions. The most notable one is that *opinion* is a transitional concept, which always reflects our attitude towards something. On the other hand, sentiments are different from opinions in that they reflect our feeling or emotion, not always directed towards something. Further still, our emotions may reflect our attitudes.

Generally speaking, the palette of human emotions is so vast, that it is hard to select even the basic ones. Most of the authors in the NLP community agree on the classification proposed by Paul Ekman and his colleagues (1982), which mentions six basic emotions: *anger, disgust, fear, joy, sadness, surprise*. Although this classification is consistent in itself, it needs to be further extended by antonyms in order to allow capturing positive and negative shifts in opinion. Accordingly, Jianwei Zhang et al. (2009) propose to group the basic emotions along four dimensions: *Joy* ⇔ *Sadness, Acceptance* ⇔ *Disgust, Anticipation* ⇔ *Surprise, and Fear* ⇔ *Anger*. However, such a division requires a rather complex processing and analysis of the input data, which is not always feasible. Therefore, the majority of the authors accept a simpler representation of sentiments according to their *polarity* (Pang and Lee, 2008):

Definition 4 (Sentiment Polarity) The polarity of a sentiment is the point on the evaluation scale that corresponds to our *positive* or *negative* evaluation of the meaning of this sentiment.

Sentiment polarity allows us to use a single dimension (rather than the four dimensions mentioned above), thus, simplifying the representation and management of the sentiment information.

3.2 Problems in Opinion Mining

In the area of Opinion Mining, studies usually follow a workflow consisting of two steps: *identify* (topics, opinionative sentences), and *classify* (sentences, documents).

In the first step, we need to identify the topics mentioned in the input data, and also associate with each topic the corresponding opinionative sentences. During this step, we may also try to distinguish between opinionative and non-opinionative phrases (i.e., perform *subjectivity identification*). This additional task is useful, because not all phrases that contain sentiment words are, in fact, opinionative. The reverse claim is also true: some of the opinionative phrases do not contain positively (or negatively) charged words. Therefore, performing this identification task can be an effective addition to the classification step in order to improve precision (Wiebe et al, 2001; Dave et al, 2003; Pang and Lee, 2004; Riloff et al, 2005; Wiebe and Riloff, 2005; Wilson et al, 2005). Furthermore, retrieval of opinionative documents evolved into a separate task with many specific algorithms, like in (Yu and Hatzivassiloglou, 2003; Ku et al, 2006; Zhang et al, 2007; He et al, 2008).

During the second step, the problem of *sentiment classification* is most often a binary classification problem, distinguishing between *positive* and *negative* texts. Nevertheless, additional classes can also be introduced, in order to make the analysis

more robust and increase the quality (i.e., granularity) of results. For example, some of the works include the *neutral* or *irrelevant* sentiment categories, which mean that there is no sentiment. By doing this, we can avoid the subjectivity identification task mentioned above, and have the classifier distinguish between opinionative and non-opinionative phrases. There is evidence that this approach has a positive effect on the precision of the final results (Koppel and Schler, 2006). Previous work Zhou and Chaovalit (2008) has also tried to improve sentiment classification by running this task separately for each of the topic's features (determined by an ontology) and averaging the output. Though, this step is generally considered as separate from topic identification (Pang and Lee, 2008).

In summary, we could argue that Opinion Mining can be viewed as a classification problem, distinguishing between several classes of sentiments (most often, *positive*, *negative* and *neutral*). This division is applicable to some extent even to the methods that produce sentiments on a numerical scale, in which case the division becomes a matter of setting thresholds (between the sentiments classes).

3.3 Development of Opinion Mining

Opinion Mining has been studied for a long time. Yet, the research in this area accelerated with the introduction of *Machine Learning* methods and the use of annotated datasets (Morinaga et al, 2002; Pang et al, 2002; Yi et al, 2003; Dave et al, 2003). Other types of approaches have also been used, like *Dictionary*, *Statistical*, and *Semantic*. Yet, since the early days of opinion mining, machine learning has been the most frequently exploited tool for tackling the relevant problems.

The **Machine Learning Approach** is a sophisticated solution to the classification problem that can be generally described as a two-step process: 1) learn the model from a corpus of training data (supervised, unsupervised), and 2) classify the unseen data based on the trained model.

Below, we provide a formal statement for the (supervised) learning step, adapted to our terminology. We assume training data are documents represented in a space, $\mathbb{D}$, whose dimensions are document features (e.g., frequency of words, bi-grams, etc.). Furthermore, these documents have been assigned a sentiment label from a space $\mathbb{S}$:

$$\text{Given training data } \{(D_i \in \mathbb{D}, S_i \in \mathbb{S})\}, \text{ find } g : \mathbb{D} \to \mathbb{S},\ g(D_i) = \arg\max_{S} f(D_i, S_i) \quad (1)$$

The above formulation says that given a set of training pairs, we want to find a function g that maps documents to sentiment labels, according to the best prediction of some scoring function f. This function takes as input documents and sentiment labels and gives a sentiment label probability prediction (using either conditional or joint probability). Without loss of generality, the learning process can be considered as an estimation of the scoring function. Examples of such scoring functions are feature vectors in $\mathbb{D}$, computed relative to class separating hyperplanes, or functions based on decision trees.

The machine learning approach involves the following general steps. First, a training dataset is obtained, which may be either annotated with sentiment labels (supervised learning), or not (unsupervised learning). Second, each document is represented as a vector of features. We describe examples of such representations further in the text. Third, a classifier is trained to distinguish among sentiment labels by analyzing the relevant features. Finally, this classifier is used to predict sentiments for new documents.

The current popularity of the machine learning approach for opinion mining originates from the work "Thumbs up?" by Pang and Lee (2002). The authors proposed and evaluated three supervised classification methods: Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM). According to their evaluation, SVM showed the best performance, while NB was the least precise out of the three (though, the differences among them were small). Nevertheless, all the algorithms clearly surpassed the random choice baseline, exhibiting an average precision of around 80%. Dave et al. (2003) further extended the work of Pang and Lee, emphasizing feature selection. They also used Laplacian smoothing for NB, which increased its accuracy to 87% (for a particular dataset). However, the SVM classifier has achieved similar results, performing below NB only when using unigram features (refer also to Table 2). Pang and Lee (2004) also used subjectivity identification as a preprocessing step in order to improve the precision of NB.

The sentiment analysis task is very 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. The need to provide a finer resolution of sentiments, without affecting the classification accuracy, required different multi-class categorization methods compared to traditional SVM. Although the SVM method has proved its efficiency for binary classification, the new problem demanded more sophisticated solutions.

To address this challenge, Pang and Lee (2005) in their study "Seeing Stars" proposed to use SVM in multi-class *one-versus-all* (OVA) and *regression* (SVR) modes, combining them with metric labeling, so that similar classes are positioned closer to each other on a rating scale. Metric labeling is a special case of *a-posteriori* optimization of class assignment with respect to *prior* assignment. This class assignment minimizes the sum of distances between labels of adjacent points, penalized by point similarities. Their results clearly demonstrated that a combination of SVM with other unsupervised classification methods results in better precision. A subsequent work on support or opposition in the context of political texts studied further extensions to the SVM approach, through modeling relationships and agreement between authors (Thomas et al, 2006).

The performance of machine learning methods is highly dependent on the quality and quantity of training data, which is scarce compared to the amount of unlabeled data. In the paper titled "Seeing Stars When There Are Not Many Stars", Goldberg and Zhu (2006) proposed a semi-supervised learning technique operating on a graph of both labeled and unlabeled data. The authors represent documents with a graph, where vertices correspond to documents, and edges are drawn between similar documents using a distance measure computed directly from document features. These assumptions are similar to metric labeling, except that they are used *a-priori*, thus, allowing to use even unlabeled data for training. Although their approach exhibited

better performance than SVR, the authors mention that it is sensitive to the choice of the similarity measure, and not able to benefit from the use of additional labeled data.

In the studies discussed above, rating inference tasks have been considered at the document level, thus showing an 'average' precision on heterogeneous reviews, which mention multiple aspects of the product with different sentiments expressed on each one. This brings up the problem of contextual sentiment classification, which requires algorithms not only operating at the sentence level, but also involving the context of each sentence in their analysis (Wilson et al, 2005). Extending on (Pang and Lee, 2005), Shimada and Endo (2008) proposed to analyze ratings on the product-feature level, naming their work "Seeing Several Stars". They have demonstrated that SVR, despite being less precise than SVM, produces output labels that are closer to the actual ones. This evidence also supports the claim in (Pang and Lee, 2005) that with the use of a "gradual" function in SVR "similar items necessarily receive similar labels".

Apart from the choice of algorithms and data selection, the performance of machine learning approaches is heavily dependent on feature selection. The most straightforward (yet, in some cases very effective) way is to encode each feature in the set by its presence or absence in the document. In the case of word features, this would produce a simple binary vector representation of a document. Extending this representation, we can instead use relative frequencies of words' occurrence. Though, not all words are equally representative and, therefore, useful for subjectivity analysis. This provides an opportunity to make the learning process more efficient by reducing the dimensionality of $\mathbb{D}$ (refer to Formula 1). Osherenko et al. (2007) demonstrate that it is possible to use just a small set of the most affective words as features, almost without any degradation in the classifier's performance. Nevertheless, the direct use of sentiment values from such dictionaries has shown little to no increase of precision. Therefore, studies usually use frequencies of words instead. For example, Devitt and Ahmad (2007) identify sentiment-bearing words in a document by using SentiWordNet, but then use just their frequencies of occurrence for the classification task. This approach is also popular with dictionary methods, which we describe below.

Finally, we should mention that machine learning is used for other problems of opinion mining, like subjectivity identification. Zhang et al. (2007) describe an approach that uses an SVM trained on a set of topic-specific articles obtained from Wikipedia (objective documents) and review sites (subjective documents).

The **Dictionary Approach** relies on a pre-built dictionary that contains opinion polarities of words, such as the General Inquirer[2], the Dictionary of Affect of Language[3], the WordNet-Affect[4], or the SentiWordNet (Esuli and Sebastiani, 2006), which is the most popular dictionary today.

Existing works exploit these resources mainly for identification of opinionative words, although some recent studies showed that it is possible to use polarity scores directly, providing a sentiment value on a continuous scale (Fahrni and Klenner, 2008; Tsytsarau et al, 2010; Missen and Boughanem, 2009). Polarity of a sentence or doc-

[2] http://www.wjh.harvard.edu/~inquirer/

[3] http://www.hdcus.com/

[4] http://wndomains.fbk.eu/wnaffect.html

ument in this case is usually determined by averaging the polarities of individual words. For instance, most of the dictionary methods aggregate the polarity values for a sentence or document, and compute the resulting sentiment using simple rule-based algorithms (Zhu et al, 2009). More sophisticated tools, like the Sentiment Analyzer introduced by Yi et al. (2003), or the Linguistic Approach by Thet et al (2009), extract sentiments precisely for some target topics using advanced methods that exploit domain-specific features, as well as opinion sentence patterns and Part-Of-Speech tags. The above two approaches lead to better performance, albeit at the expense of increased computational complexity.

We now describe a formula that defines the most general case of document opinion assignment using a dictionary:

$$S(D) = \frac{\sum_{w \in D} S_w \cdot weight(w) \cdot modifier(w)}{\sum weight(w)} \tag{2}$$

In the above equation, S_w represents the dictionary sentiment for a document word w, which is being aggregated with respect to some weighting function $weight()$ and modifier operator $modifier()$ (which handles negation, intensity words, and other cases affecting *a-priori* sentiment). Weighting functions may be defined statically for each sentence, or computed dynamically, with respect to positions of topic words. Usually weighting functions represent a window around the topic word, thus, taking into account the sentiments of the words that are immediate neighbors of the topic word in the document. For example, a weighting function can have the value of 1 for two or three words surrounding the topic word, and 0 elsewhere. More sophisticated methods may also be used, such as NLP processing, which can lead to a dynamic computation of the weighting function for each sentence, taking into account its specific structure.

Nevertheless, the use of dictionaries can also be combined with machine learning methods, as we mention in the previous paragraphs. We note that relying on the polarity values assigned by a dictionary is not always feasible, as the dictionary may not be suited for use on particular datasets (e.g., may not include some domain-specific lexicons). Furthermore, dictionary methods are usually not able to adapt polarity values to particular contexts. It turns out that words can change their polarity when used in different contexts (Fahrni and Klenner, 2008). Consider the adjectives “cold” (generally regarded as negative), and “warm” (regarded as positive). When these adjectives are used in the phrases “cold wine” and “warm beer”, their polarities change to positive and negative, respectively. In contrast to the dictionary approach, machine learning methods naturally adapt to the corpus they are trained on.

The **Statistical Approach** aims to overcome the problems mentioned above. For example, Farni and Klenner (2008) propose to derive posterior polarities using the co-occurrence of adjectives in a corpus. In this case, adaptability is achieved through the construction of a corpus-specific dictionary. Regarding the problem of unavailability of some words, the corpus statistics method proposes to overcome it by using a corpus that is large enough. For this purpose, it is possible to use the entire set of indexed documents on the Web as the corpus for the dictionary construction (Turney, 2002).

We can identify the polarity of a word by studying the frequencies with which this word occurs in a large annotated corpus of texts (Leung et al, 2006; Miao et al, 2009). If the word occurs more frequently among positive (negative) texts, then it has a positive (negative) polarity. Equal frequencies indicate neutral words. It is also interesting to mention, that applications working with the Chinese language are able to recognize polarity even for unseen words, thanks to the fact that phonetic characters determine the word's sense (Ku et al, 2006, 2007). In this case, we can analyze frequencies of single characters rather than words. Although computationally efficient, the basic method requires a large annotated corpus, which becomes a limiting factor.

The state of the art methods are based on the observation that similar opinion words frequently appear together in a corpus. Correspondingly, if two words frequently appear together within the same context, they are likely to share the same polarity. Therefore the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word, which invariantly preserves its polarity (an example of such a word is "good"). To achieve this, Peter Turney (2002; 2003) proposed to use the Point-wise Mutual Information (PMI) criterion for statistical dependence (Church and Hanks, 1989), replacing probability values with the frequencies of term occurrence $F(x)$ and co-occurrence $F(x \textit{ near } y)$:

$$PMI(x,y) = \log_2 \frac{F(x \textit{ near } y)}{F(x)F(y)}; \tag{3}$$

Sentiment polarity (expressed by *PMI-IR*) for word x is then calculated as the difference between *PMI* values computed against two opposing lists of words: positive words, *pWords*, such as "excellent", and negative words, *nWords*, such as "poor":

$$PMI\text{-}IR(x) = \sum_{p \in pWords} PMI(x,p) - \sum_{n \in nWords} PMI(x,n) \tag{4}$$

Along with the formulas above, Turney et al. proposed to obtain the co-occurrence frequencies F by relying on the statistics of the AltaVista web search engine. Extending on this work, Chaovalit et al. (2005) used Google's search engine to determine the co-occurrence of words, increasing the precision. Read et al. (2009) further extended this approach, employing Semantic Spaces and Distributional Similarity as alternative weakly-supervised methods. A detailed study on constructing dictionaries of this kind was made by Taboada et al. (2006), mentioning some problems that occur due to the unavailability of the "near" modifier or non-persistence of the search engine's output. On the other hand, search engines allow retrieving the co-occurrence scores (thus, polarities) not only for words, but also for phrases, which is a useful feature.

The use of statistical methods in computing opinion polarity has found an interesting development in the work of Ben He et al. (2008), where they propose to use an opinion dictionary along with IR methods in order to retrieve opinionative blog posts. Their approach first builds a dictionary by extracting frequent terms from the entire collection, which are then ranked according to their frequency among opinion-annotated texts. The sentiment polarity of each document is computed as a relevance score to a query composed of the top terms from this dictionary. Finally, the opinion relevance score is combined with the topic relevance score, providing a ranking of opinionative documents on that topic.

The **Semantic Approach** provides sentiment values directly (like the Statistical Approach), except that it relies on different principles for computing the similarity between words. The underlying principle of all approaches in this category is that semantically close words should receive similar sentiment values.

WordNet (Fellbaum, 1998) provides different kinds of semantic relationships between words, which may be used to calculate sentiment polarities. The possibility to disambiguate senses of words using WordNet can serve as a way to include the context of these words into the opinion analysis task. Similar to statistical methods, two sets of seed words with positive and negative sentiments are used as a starting point for bootstrapping the construction of a dictionary.

Kamps et al. (2004) proposed to use the relative shortest path distance of the "synonym" relation, demonstrating a good degree of agreement (70%) with an annotated dictionary. Another popular way of using WordNet is to obtain a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms (Kim and Hovy, 2004; Hu and Liu, 2004a). The sentiment polarity for an unknown word is determined by the relative count of positive and negative synonyms of this word (Kim and Hovy, 2004). Otherwise, unknown words may also be discarded (Hu and Liu, 2004a). However, it is important to know that since the synonym's relevance decreases with the length of the path between the synonym and the original word, so should the polarity value, too. Additionally, the polarity of a word is often averaged over all possible paths to it. Though, as was pointed out by Godbole et al. (2007), we should only consider paths that go through the words of the same polarity as initial.

3.4 Opinion Mining in Microblogs and Streaming Data

In the above paragraphs, we mostly considered static approaches to the problem of Opinion Mining, where the classifier's model does not change after being constructed. However, there exists another class of applications, such as those analyzing messages in microblogging, which require adaptability of the model to changing data during the analysis.

The most prominent example of microblogging platforms, which allows for real-time analysis, is Twitter. Its vast user community, all-around presence and informal style of conversation make Twitter a rich source of up-to-date information on different events and a good indicator of users' moods. Recently, it was demonstrated that sentiments from Twitter messages correlate with political preferences (Tumasjan et al, 2010), and even improve stock market prediction (Bollen et al, 2010).

Recent works have identified several differences between opinion mining in microblogs when compared to conventional opinion analysis of documents. The main difference is the availability of sentiment or mood annotations in messages, providing a good source of training data for classifiers (Go et al, 2009; Bifet and Frank, 2010; Pak and Paroubek, 2010).

Pak and Paroubek (2010) performed statistical analysis of linguistic features of Twitter messages and report interesting patterns which may help distinguish among sentiment classes. They demonstrate that an NB classifier, based on negation extended bi-gram features, achieves good accuracy (albeit, at the expense of low recall)

and can be useful to information retrieval applications. Bermingham and Smeaton (2010) compared the performance of SVM and Multinomial Naive Bayes (MNB) classifiers on microblog data and reviews, and demonstrated that in most cases these classifiers yield better results on short-length, opinion-rich microblog messages.

Since class distributions may vary along the stream of data, there is a necessity to follow these changes and update the classifier's model accordingly. Bifet and Frank (2010) studied the problem of using an adaptable classifier with Twitter data and examined relevant evaluation methods. They proposed to use the Stochastic Gradient Descent (SGD) method to learn a linear classifier. The proposed approach allows specifying the rate with which model's parameters are updated, and to monitor the evolution of the impact individual words have on class predictions. The latter may be used as an indicator of users' support or opposition to particular topics in a stream. In addition, SGD demonstrated an accuracy smaller but comparable to that of MNB (67.41% versus 73.81%).

4 Opinion Aggregation

The analysis of opinions at a large scale is impractical without automatic aggregation and summarization. In this case, we are interested in identifying opinions at a higher level than that of an individual: we would like to identify the average or prevalent opinion of a group of people about some topic, and track its evolution over time.

What distinguishes Opinion Aggregation from other tasks, is the necessity to provide summaries along several features, aggregated over one or more dimensions. Therefore, feature extraction and aggregation appear as the key problems here, and we are going to concentrate our attention on these tasks.

The problem of mining product reviews has attracted particular attention in the research community (Morinaga et al, 2002; Dave et al, 2003; Liu et al, 2005; Carenini et al, 2005). This problem imposes certain challenges related to the extraction of representative features and the calculation of the average sentiment or rating. The final goal though, is to determine the overall opinion of the community on some specific product, rather than the individual user opinion on that product.

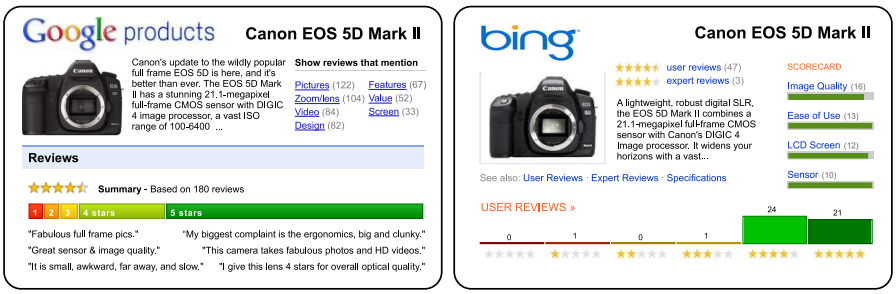

Fig. 1: An example of Google and Bing review aggregations (actual images and text were arranged for better representation).

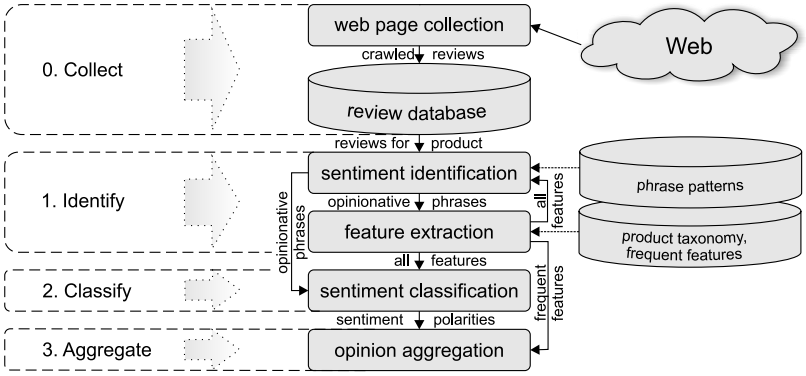

Fig. 2: An example architecture of product review aggregation.

Today we can already see working examples of opinion aggregation at several web sites that visualize collaborative ratings assigned by a community of users. In Figure 1, we depict two examples of opinion aggregation, from the Google and Bing web search engines. Both of them feature images, short descriptions, and aggregate ratings. Additionally, they include statistics for each rating category (number of "stars"). Overall, these two approaches show similar details on the featured product, except that Google offers a representative summary (sentences at the bottom), while Bing displays aggregated ratings for each product feature (displayed on the right).

4.1 Problems in Opinion Aggregation

Review mining is the main application domain for opinion aggregation. Therefore, the problems that have been studied in relation to opinion aggregation are mainly formulated around the aggregation of product reviews. They include the processes of collecting, mining and reasoning on customer feedback data, represented in the form of textual reviews (Tang et al, 2009).

Figure 2 illustrates the review mining process. The process starts with the identification of opinionative phrases, which may additionally involve a collection of phrase patterns, or comparative sentences (in this case, sentiments are expressed by means of comparison of an object to another similar object) (Liu, 2010). Identified phrases are then passed on to the feature extraction step, which may exploit a product taxonomy database (Carenini et al, 2005) in order to improve the results. Features and opinionative phrases are used in the sentiment classification step, which outputs sentiment polarities to be aggregated over frequent features at the opinion aggregation step. This process can be iterative, using the identified features in order to improve the phrase extraction step.

Although Opinion Aggregation is a separate task having its own problems, practical applications also involve information retrieval and sentiment analysis techniques during the data pre-processing. Thus, the Opinion Aggregation techniques have been developing in close connection to other methods, and were subsequently revisited

when improved sentiment analysis and feature extraction methods were introduced. Generally speaking, Opinion Aggregation methods are quite modular and may be used with different Opinion Mining algorithms. For example, Carenini et al. (2005) describe a system that relies on sentiment extraction only as a preprocessing task, concentrating their attention on the aggregation of user reviews.

Aggregation of opinions for a product, expressed in a document collection $\mathscr{D}$, may be formulated as the problem of determining a set of product features (each labeled with a corresponding sentiment), satisfying certain criteria:

$$\{(f, \mu_S)\} \mid rep(f, \mathscr{D}) > \rho_f,\ \mu_S = agg(S, f),\ \text{satisfying}\ con(S) \tag{5}$$

Where f is a product feature that is important for the description of the product in $\mathscr{D}$, according to some representativeness measure $rep()$, and μ_S is the sentiment for f, computed over $\mathscr{D}$ according to some aggregating function $agg()$. During this procedure, we may only consider features with a representativeness measure over some threshold ρ_f, and corresponding sentiments that satisfy some constraints, expressed by $con(S)$. Examples of such constraints are imposing a limit on the sentiment's absolute value (e.g., consider only moderate opinions), or the timestamp (e.g., consider only recent opinions).

We note that Opinion Aggregation is different from (text) summarization, which is the problem of producing a shortened version of the corresponding text. These problems are complementary to each other, and in this study we focus on the former since it involves Opinion Mining.

4.2 Development of Opinion Aggregation

A typical method for opinion aggregation was proposed by Hu et al. (2004a). They describe a system that aims at discovering words, phrases, and sentiments that best characterize some product. At a high level, their solution follows the steps we listed in the previous section. We note though, that not all studies follow this pattern. For example, Morinaga et al. (2002) reversed the ordering of steps 1 and 2, and the experiments revealed that their system achieves a similar performance. By running opinion classification prior to identification of features, we effectively apply some kind of filtering on features: we remove those that were not mentioned in an opinionative phrase (since these are features that are irrelevant for our analysis).

Different approaches to feature extraction have been proposed. Hu et al. (2004b) identify features by building a list of noun-noun phrases using an NLP parser, and then determining the most frequent ones. Feature frequency in this case corresponds to the $rep()$ function in Formula 5. However, their approach outputs many irrelevant words and should be used in conjunction with other methods, as was suggested by Carenini et al. (2005). Accordingly, they introduce a domain taxonomy in the form of user-defined features, which are used to annotate data for training a feature classifier. Opinions are then collected and aggregated based on the full set of features, which consists of features extracted automatically (unsupervised learning) and also through the classifier (supervised learning). Alternatively, Ku et. al. (2006) proposed a system that identifies features by using information retrieval methods. They use TF-IDF

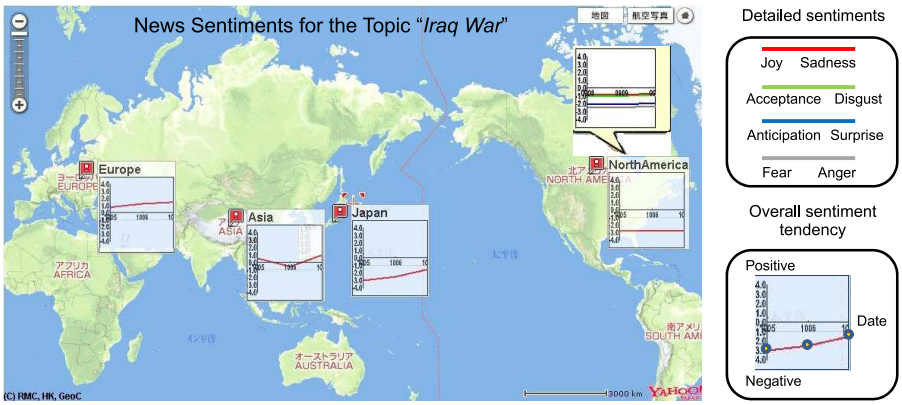

Fig. 3: An example of geographical sentiment aggregation (Zhang et al, 2009).

scores per paragraph and per document, and a dictionary to determine polarity. The intuition here is that relevant features appear frequently in few of the paragraphs of many documents, or in many of the paragraphs of few documents. This technique is also efficient for eliminating the irrelevant features described above.

Aggregation of opinions has been traditionally performed over all the documents in some collection. Miao et al. (2009) proposed a time-decaying aggregation approach, retrieving only the most recent reviews that were marked by users as helpful. The above constraints are represented by the *con*() function in Formula 5. Jianwei Zhang et al. (2009) introduced a novel technique, which interactively aggregates and displays sentiments based on different granularities of time and space (geographical location). Moreover, the sentiments are represented by several dimensions, making it the most robust Web-scale application we observed in our study. An example of such an aggregation is shown in Figure 3. In this figure, we can see a world map and the time evolution of the sentiments in news articles for different geographical regions of the world. These results are illustrated in pop-up boxes, which report the values of four sentiment dimensions (i.e., joy-sadness, acceptance-disgust, anticipation-surprise, and fear-anger) over time. This system automatically retrieves and displays sentiments around some particular time period for *ad-hoc* queries, aggregating them over different locations as the user navigates the map, or zooms in and out.

4.3 Opinion Quality and Spam

With the rapid growth of web sites featuring product ratings and their increasing impact on users' opinions and/or buying decisions, it comes as no surprise that we observe a significant interest to this area from commercial organizations (Hoffman, 2008). These organizations include product manufacturers, marketing and advertising agencies. Opinion Aggregation plays an important role in this field, because it has the potential to capture the opinions of the community. However, it also has some

weak points, such as the smoothing of the variances in opinions and the possibility to manipulate the aggregate values by introducing artificially constructed data. This makes opinion quality assessment and spam detection useful pre-processing steps for Opinion Aggregation.

The first problem, opinion quality assessment, aims at determining the quality of opinions expressed in a review. Lu et al (2010) describe an opinion quality classifier relying not only on the review's textual features, but on the reviewer's social context, as well. The authors propose a method that optimizes an error function for training data in a feature space, subject to four regularization constraints. These constraints capture the intuition that the quality of reviews from the same user, as well as from users connected in a social context to that one, should be about the same. The introduced constraints do not employ annotated labels, therefore, may be used to train a model on unlabeled data. The study shows that the proposed method increases the accuracy of identifying reviews of high quality.

At the same time, we observe that the phenomenon of opinion spam (or fake reviews) is also growing (Jindal and Liu, 2008; Chen et al, 2009; Lim et al, 2010). The detection of opinion spam is a hard problem, since spam is targeted to specific products (therefore, resistant to aggregation), and not easily distinguishable from real reviews. This problem had not been studied in depth until recently. Below, we briefly discuss two of the papers in this area that are relevant to Subjectivity Analysis. The aim of these studies is to identify opinion spam in a pre-processing step. Then, the review spam can be excluded from further consideration, thus, resulting in more accurate and truthful Opinion Aggregation.

The work of Lim et al. (2010) proposes a method for detecting spammers, rather than individual spam reviews. Each user is attributed with the following statistical measures: *Rating Spamming*, which is the average similarity among the user's ratings for each product; *Review Text Spamming*, which is the average similarity among the user's review texts for each product; *Single Product Group Multiple High (Low) Ratings*, which is the number of extremely high (low) ratings posted by the user for each product group in time intervals where such quantity exceeds a certain threshold; *General Deviation*, which is the average deviation of the user's ratings from the mean of each product; *Early Deviation*, which is the same as General Deviation, only weighted according to time. All the individual measures are normalized against all users, and the overall final measure for each user is computed as their weighted sum. To classify a user as spammer, one needs to compare the user's measure against some threshold.

Jindal and Liu (2008) classify opinion spam into the following three categories: *untruthful opinions*, *reviews on brands only* and *non-reviews*. The first category, *untruthful opinions*, is represented by intentionally biased reviews, either positive or negative. The second category, *reviews on brands only*, consists of opinions about some brand in general, without discussion of specific product features. The third category, *non-reviews*, refers to explicit advertisement, technical data or off-topic text. The authors propose classification techniques to identify the spam reviews, the ones belonging to the first category being the most challenging to discover.

5 Contradiction Analysis

By analyzing a community's opinions on some topic, we understand how people in general regard this topic. However, people do not always share the same opinions on different topics. Therefore, opinion aggregation may produce a lossy summarization of the available opinion data, by ignoring and masking the diversity that inherently exists in data. In order to find an answer to this interesting problem, we have to employ more advanced techniques, as we discuss in this section.

In several cases, performing simple aggregations on opinions is not enough for satisfying the requirements of modern applications. We may be interested in focusing on the topics for which conflicting opinions have been expressed, in understanding these conflicting opinions, and in analyzing their evolution over time and space. Evidently, we need to be able to effectively combine diverse opinions in *ad hoc* summaries, and also to further operate on these summaries in order to support more complex queries on the dynamics of the conflicting, or contradicting opinions. An example of a problem requiring this kind of complex analytics is *Contradiction Analysis*, an emerging research direction under the general area of Subjectivity Analysis.

5.1 Definitions of Contradiction Analysis

The contradiction analysis area is a relatively new direction of research. As such, there is no established common framework for describing and modeling the relevant problems. Though, some recent studies have made the first steps towards this direction.

De Marneffe et al. (2008) introduce a classification of contradictions consisting of seven types that are distinguished by the features that contribute to a contradiction (e.g., antonymy, negation, numeric mismatches). Antonymy are words that have opposite meanings, i.e., "hot - cold" or "light - dark". Antonymy can give rise to a contradiction when people use these words to describe some topic. Negation imposes a strict and explicit contradiction, e.g., "I love you - I love you not". Numeric mismatches form another type of contradiction, which may be caused by erroneous data: "the solar system has 8 planets - there are 9 planets orbiting the sun". Their work defines contradictions as a situation where "two sentences are extremely unlikely to be true when considered together". In other words, contradictions may be defined as a form of textual entailment, when two sentences express mutually exclusive information on the same subject (Harabagiu et al, 2006).

The works discussed above rely on human-perceivable definitions of contradiction that summarize our expectations about which features contribute to a contradiction. Opposite sentiments are also very common sources of contradictions. However, they may be described in different terms compared to the textual entailment problem. Consider the following example: "I like this book - This reading makes me sick". Both sentences convey a contradiction on opinions expressed about a book, yet they may appear together if they belong to different authors. Therefore, we may relax the 'exclusivity' constraint of textual entailment and propose the following definition:

Definition 5 (Contradiction) There is a contradiction on a topic, T, between two sets of documents, $\mathscr{D}_1, \mathscr{D}_2 \subset \mathscr{D}$ in a document collection $\mathscr{D}$, where $\mathscr{D}_1 \bigcap \mathscr{D}_2 = \emptyset$, when the information conveyed about T is considerably more different between $\mathscr{D}_1$ and $\mathscr{D}_2$ than within each one of them.

In the above definition, we purposely not specify what it means to have some information on a topic to be very different from another piece of information (on the same topic). This definition captures the essence of contradictions, without trying to impose any of the different interpretations of what might cause a contradiction to arise. For example, if we assume that opinion polarity is the relevant information, then a contradiction would mean that two groups of documents express contrasting opinions on some topic.

When identifying contradictions in a document collection, it is important to also take into account the time in which these documents were published. Let $\mathscr{D}_1$ be a group of documents containing some information on topic T, and all documents in $\mathscr{D}_1$ were published within some time interval t_1. Assume that t_1 is followed by time interval t_2, and the documents published in t_2, $\mathscr{D}_2$, contain a conflicting piece of information on T. In this case, we have a special type of contradiction, which we call *Asynchronous Contradiction*, since $\mathscr{D}_1$ and $\mathscr{D}_2$ correspond to two different time intervals. Following the same line of thought, we say that we have a *Synchronous Contradiction* when both $\mathscr{D}_1$ and $\mathscr{D}_2$ correspond to a single time interval, t.

An interesting application of contradiction analysis is in supplementing information retrieval systems, which in most of the cases are fact-centric. Diverse opinions introduce extra noise to such systems, which are intended to provide a solid and unbiased representation of information about different topics (Riloff et al, 2005). Understanding contradicting opinions allows information retrieval systems to deal with opinionative data using special methods, for example by extracting the ground truth from different discussions or representing user support against different conflicting topics.

5.2 Problems in Contradiction Analysis

A typical Contradiction Analysis application needs to follow the same steps we identified for Opinion Mining, namely, topic identification and sentiment extraction. For certain techniques of Contradiction Analysis it is possible to rely directly on the output of Opinion Mining, thus simplifying the entire workflow. Then, we need to have a contradiction detection step, where individual sentiments are processed in order to reveal contradictions.

In the contradiction detection step, the goal is to efficiently combine the information extracted in the previous steps, in order to determine the topics and time intervals in which contradictions occur. In this step, statistical methods can be used, as well as clustering, or other unsupervised methods. The contradiction detection step requires efficient data mining methods, which will enable the online identification of contradictions, and will have the ability to work on different time resolutions.

5.3 Development of Contradiction Analysis

As with all other Subjectivity Analysis problems, research on Contradiction Analysis is under way in different domains. It is interesting to mention that the identification of contradicting claims first appeared in the speech recognition domain. The works by Hillard et al. (2003) and Galley et al. (2004) established it as a problem of recognizing agreement (positive) and disagreement (negative) texts, by looking at sentiments and negation. The authors exploited machine learning techniques for classification purposes, combining audio and text features.

Another approach to contradiction detection is to handle it as a textual entailment problem. There are two main approaches, where contradictions are defined as a form of textual inference (e.g., entailment identification) and analyzed using linguistic technologies. Harabagiu et al. (2006) present a framework for contradiction analysis that exploits linguistic information (e.g., types of verbs), as well as semantic information, such as negation or antonymy. Further improving the work in this direction, de Marneffe et al. (2008) define several linguistic features that contribute to a contradiction (discussed in Section 5.1). Exploiting these features, supplemented by the sentence alignment tool, they introduced a contradiction detection approach to their textual entailment application (Pado et al, 2008).

Although the detection of contradictions using linguistic analysis and textual entailment promises more accurate results overall, the current methods do not yet achieve high precision and recall values (Voorhees, 2008; Giampiccolo et al, 2008). For example, Pado et al. (2008) report their precision and recall values of contradiction detection at the RTE-4 task as being 28% and 8%, respectively. Therefore, scientists concentrate their efforts in finding contradictions of only a specific type when dealing with large-scale web analysis. In particular, they analyze negation and opposite sentiments.

Ennals et al. (2010a; 2010b) describe an approach that detects contradicting claims by checking whether some particular claim entails (i.e., has the same sense as) one of those that are known to be disputed. For this purpose, they have aggregated disputed claims from Snopes.com and Politifact.com into a database. Additionally, they have included disputed claims from the web, by looking for an explicit statement of contradiction or negation in the text. Although this approach would not reveal all types of contradictions, it can help to identify some obvious cases, which can be further used as seed examples to a bootstrapping algorithm.

The problem of identifying and analyzing contradictions has also been studied in the context of social networks and blogs. Relying on the exploited data mining algorithms, scientists proposed different measures for contradiction. Choudhury et al. (2008) examine how communities in the blogosphere transit between high- and low-entropy states across time, incorporating sentiment extraction. According to their study, entropy grows when diversity in opinions grows. A recent work by Liu et al. (2009) introduces a system that allows comparing contrasting opinions of experienced blog users on some topic. Then, they aggregate opinions over different aspects of the topic, which improves the quality and informativeness of the search results.

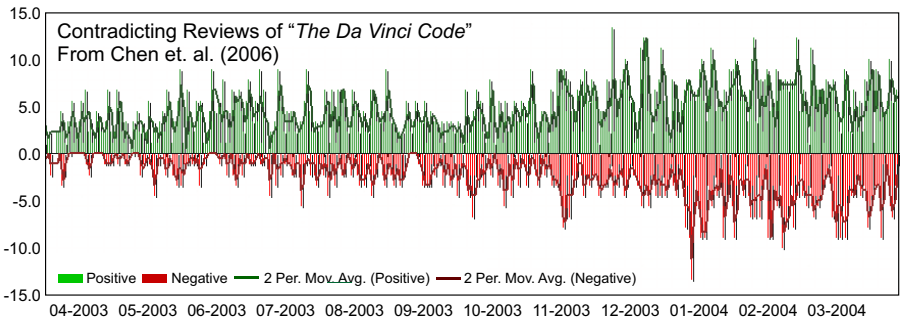

Fig. 4: Opinion timeline visualization (Chen et al, 2006).

In some cases it is also interesting to examine how the blog entries of a single user change over time. The study in (McArthur, 2008) focuses on the analysis of the sentiments of individual users, and how these change as a function of time. Similar to the approaches we discussed in the previous paragraph, the contradicting opinions are not aggregated. It is up to the user to visually inspect the results and draw some conclusions.

Kim and Zhai (2009) also propose a novel contrastive opinion summarization problem, which aims at extracting representative, but diverse, opinions from product reviews (annotated with sentiment labels during preprocessing). Their solution is based on the measures of *representativeness r* and *contrastiveness c*.

$$r = \frac{1}{|X|}\sum_{x\in X}\max_{i\in[1,k]}\phi(x,u_i) + \frac{1}{|Y|}\sum_{y\in Y}\max_{i\in[1,k]}\phi(y,v_i), \qquad c = \frac{1}{k}\sum_{i=1}^{k}\psi(u_i,v_i) \tag{6}$$

The first measure is based on the weighted sums of maximal content similarities, ϕ, among positive, X, and negative, Y, sets of sentences and their corresponding summaries, u and v. Representativeness reflects how well the summaries approximate the original text. Contrastiveness captures the similarity between positive and negative sentences in the summaries, but is computed based on the contrastive similarity ψ that is the same as content similarity, except that it is computed without taking into account sentimental words. Elimination of sentimental words results to improved precision for this similarity matching. Both ϕ and ψ rely on similarities among a review's individual words, either restricted to an exact match or a semantic (probabilistic) match.

Chen et al. (2006) study the problem of contradicting opinions in a corpus of book reviews, which they classify as positive and negative. The main goal of their work is to identify the most predictive terms for the above classification task, but the results are also used to visualize the contradicting opinions. An example of such visualization can be seen in Figure 4. The visualization is composed by two trends of opposite (positive, negative) opinions, along with their moving averages. The user can determine contradicting regions by visually comparing these trends. However, such an analysis, which is based on manual inspection, does not scale and becomes cumbersome and error-prone for large datasets.

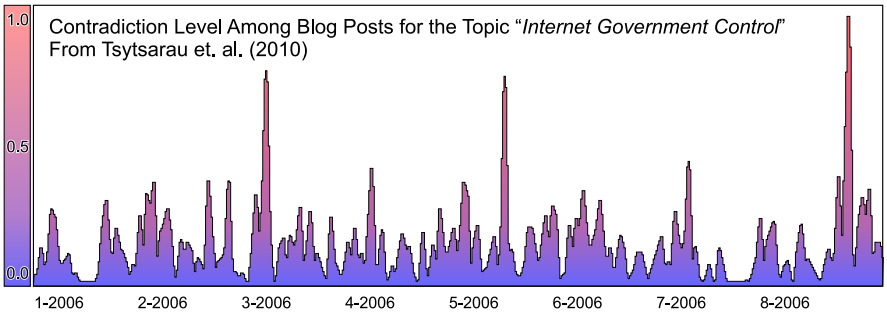

Fig. 5: Contradiction timeline visualization (Tsytsarau et al, 2010).

Tsytsarau et al. (2010; 2011) propose an automatic and scalable solution for the contradiction detection problem. In their work, they study the contradiction problem by focusing on the analysis of sentiments. An example result of such an analysis is represented in Figure 5, which depicts the evolution of the contradiction level for the topic "internet government control", covering a time period of about one year. The graph shows the peaks in contradiction for this topic, enabling the analyst to focus on the interesting time points (and the corresponding documents) along the time interval.

The intuition behind the proposed contradiction measure is that when the aggregated value for sentiments (on a specific topic and time interval) is close to zero, while the sentiment diversity is high, then the contradiction should be high. The authors define the *Aggregated Sentiment* μ_S as the mean value over all individual sentiments, and *Sentiment Diversity* σ_S^2 as their variance. Combining μ_S and σ_S^2 in a single formula, the authors propose the following measure for contradictions:

$$C = \frac{\vartheta \cdot \sigma_S^2}{\vartheta + (\mu_S)^2} W, \qquad W = [1 + exp(\frac{\overline{n} - n}{\beta})]^{-1} \tag{7}$$

where n is the cardinality of $\mathscr{D}$, and W is a standard weight function to account for the varying number of sentiments in the time interval of interest. The constant $\overline{n}$ is the expected arrival rate (i.e. average number) of sentiments, and β is the deviation of the rate; these parameters can be estimated based on past data. The factor $\vartheta \neq 0$ limits the level of contradiction C when $(\mu_S)^2$ is close to zero.

Contradictions may occur not only on the opinion level, but also on the topic level. For example, Varlamis et al. (2008) propose clustering accuracy as an indicator of the blogosphere topic convergence. Clustering accuracy (when represented by the utility function) measures the relative separation of the cluster centers with respect to cluster sizes and a number of unclustered blogs (noise). When the clustering is very good, this function reaches its maximum value. It is easy to demonstrate, that divergence in topics leads to greater separation of individual blogs in the feature space and, therefore, less reliable clustering. By analyzing how accurate the clustering is in different time intervals, one can estimate how correlated or diverse the blog entries are. We note that this approach is relevant to the contradiction definition we gave earlier, in the sense that clustering is often defined as the process of finding distant (i.e., contradicting) groups of similar (i.e., non-contradicting) items. However, the type of contradictions that this approach discovers depends on the selection of features.

6 Discussion

In this section, we elaborate on the emerging trends, compare the various methods that have been proposed for Subjectivity Analysis, and list open problems and interesting future research directions.

6.1 Analysis of Trends

We now discuss some trends that emerge when analyzing the recent publications on Opinion Mining (for a complete list of these papers, refer to Table 1).

We allocate the papers to several classes under different dimensions: based on the employed algorithms, datasets used for testing, and target domains. In Table 1 we list several of the properties of the papers we used for the above analysis, providing a more detailed view of these studies. Here, opinion classification and opinion aggregation types are denoted by *C* and *A* correspondingly. Column *"Topic"* lists whether algorithm uses topic-specific features, linguistic parsing, domain knowledge or other techniques that allow topic-dependent analysis. Column *"Range"* lists number of the resulting sentiment categories, or *C* in the case of continuous range. Column *"Scope"* represents target domains (and subdomains) for each algorithm, which were either explicitly mentioned by the authors, or inferred from the training and testing data used in the corresponding papers (*M* - movies, *P* - products, *B* - books, *S* - various services, e.g. restaurants and travels, *A* - all or indifferent; we note that for some of the papers we reviewed this detailed information is missing). Column *"Data"* lists one or more used datasets, which are listed in Table 3. Finally, column *"Scale"* represents a characterization of the algorithm (*S* - small, *M* - medium, *L* - large) with respect to its performance and adaptability as follows. Specialized algorithms, or algorithms with high complexity (e.g., sophisticated NLP tools) were classified as small scale. Algorithms, featuring moderate performance were assigned to medium scale. Finally, we classified as large scale those algorithms that are scalable, work in an unsupervised way or may incrementally adapt as they process more data. We note that, even though this classification may not be absolutely objective, it is still useful in order to reveal some interesting trends.

In Figure 6, we depict the distribution of papers (using stacked bars) along the most popular types of algorithms and sentiment representations. We observe that the majority of the publications use machine learning methods as the classification tool of choice. Next to them are the dictionary-based methods. Under this category, we also include corpus statistics and semantic approaches. Hybrid methods that combine the above approaches (usually a combination of dictionary methods with NLP tools), are not that popular yet, probably due to their high complexity.

Regarding the representation of sentiments, the alternative approaches are to use a binary representation (i.e., two classes, positive and negative), discrete (i.e., more than two classes; the algorithms we examined used up to six), or continuous (i.e., sentiments represented using scalar values) (refer to Figure 6). Most of the approaches in the literature use the binary representation. Though, the other two representations have recently gained in popularity, since they offer finer resolution and level of control. The relatively low amount of studies featuring the discrete sentiment represen-

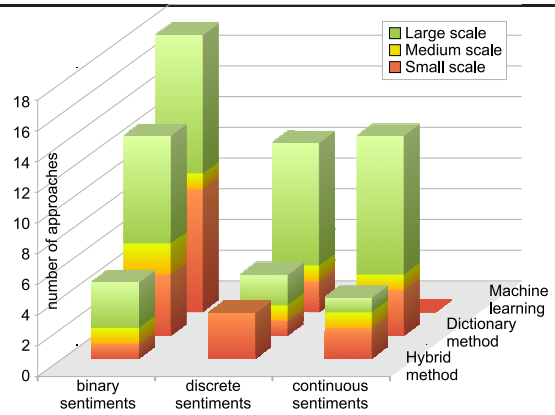

Fig. 6: The number of algorithms (stacked bars) according to sentiment representation, algorithmic approach, and scalability of the method.

tation for hybrid and dictionary methods can be explained by the availability of the continuous sentiment representation, which offers better precision. These studies use either the binary or the continuous representations, depending on their purpose. On the other hand, the continuous representation is not favored by the classification algorithms, making it a rare choice for the machine learning approaches.

The colors in each bar in the graph correspond to the number of algorithms capable of working with large, medium and small-scale datasets (green, yellow, and red color, respectively). This is directly related to the complexity of the proposed algorithms (e.g., there exist algorithms that operate only in a supervised mode, and evidently cannot scale with the dataset size). The graph shows that there are mainly two approaches that favor large-scale operation, namely, dictionary methods on continuous scale, and machine learning methods with binary and discrete representations. However, their popularity comes from different sources. Dictionary methods have the ability of unsupervised rule-based classification, which is simple and computationally efficient. On the other hand, machine learning methods achieve superior results and domain adaptability by paying the cost of the training phase. Nevertheless, they remain competitive in terms of computational complexity for the inference task (after the classifier has been constructed).

Figures 7 and 8 show the evolution of the scalability of the approaches proposed in the literature over the last years, as well as the application domains on which these approaches focused. We observe that at the beginning the majority of the studies analyzed review data, mostly at a large scale. As we mentioned above, the machine learning tools were the main contributors to this trend. The use of NLP methods since 2006 opened a new trend of complex review analysis, yet only on small scale datasets, due to the computational complexity of these methods. At approximately the same time, another interesting pattern emerged, namely, the analysis of news and social media. The current trend shows that social networks and online sources of information are attracting increasingly more interest in the research community.

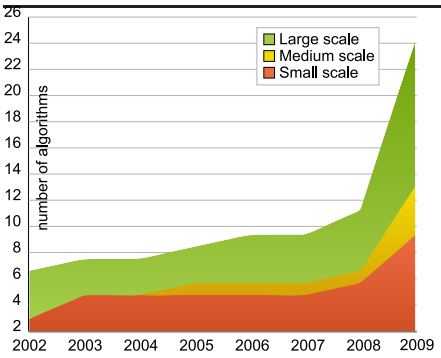

Fig. 7: Number of algorithms with different scalability levels over the last years.

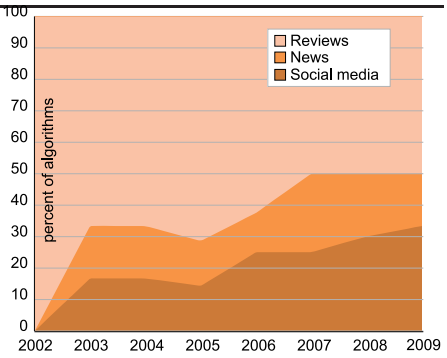

Fig. 8: Percentage of algorithms targeting different domains over the last years.

Year	Authors	Type	Topic	Algorithms	Range	Scope	Data	Scale
'02	Morinaga et al	A	Y	RuleBased+Dictionary	C	Reviews (P)	N/A	L
'02	Turney	C	N	Statistic	C	Reviews(M,P,S)	EP	L
'02	Pang et al	C	N	NB, ME, SVM	2	Reviews (M)	IMDB	L
'03	Liu et al	C	N	NLP + Dictionary	C	Texts	N/A	S
'03	Turney and Littman	C	N	LSA, Statistic (PMI)	C	Words	GI, HM	S
'03	Dave et al	A	N	NB	2	Reviews (P)	AZ, CN	L
'03	Yi et al	C	Y	Dictionary	3	Reviews (M,P)	N/A	L
'03	Yu and Hatzivassiloglou	C	N	Statistic	3	News	TREC	L
'04	Kim and Hovy	C	Y	Semantic	2	News	DUC	S
'04	Galley et al	C	N	ME, CMM	2	Transcripts	N/A	S
'04	Hu and Liu	A	Y	Semantic + RuleBased	2	Reviews (P)	AZ, CN	L
'04	Gamon	C	N	SVM	4	Reviews(S)	N/A	L
'04	Kamps et al	C	N	Semantic	C	Texts	GI	S
'05	Alm et al	C	N	Linear Classifier	2	Fairytales	N/A	S
'05	Ku et al	A	Y	Dictionary	2	News	TREC	L
'05	Chaovalit and Zhou	C	N	ML, Statistic (PMI)	2, C	Reviews (M)	IMDB	L
'05	Liu et al	A	Y	Semantic + RuleBased	2	Reviews (P)	N/A	M
'05	Pang and Lee	C	N	SVM OVA, SVR + ML	3, 4	Reviews (M)	IMDB	S
'06	Thomas et al	C	N	Multi SVM	2	Transcripts	GovTrack	S
'06	Leung et al	C	N	Statistic	3	Reviews (M)	N/A	S
'06	Taboada et al	C	N	Statistic (PMI)	2	Reviews	EP	L
'06	Carenini et al	A	Y	Semantic	2	Reviews (P)	N/A	M
'06	Ku et al	A	Y	Statistic	C	News, Blogs	TREC, NTCIR	L
'06	Goldberg and Zhu	C	N	Graph, SVR	4	Reviews (M)	IMDB	L
'06	Taboada et al	C	N	Dictionary	C	Reviews (B)	N/A	S
'07	Godbole et al	C	Y	Semantic	C	News, Blogs	N/A	L
'07	Osherenko and André	C	N	SVM + Dictionary	4	Texts	SAL	L
'07	Zhang et al	C	Y	SVM	2	Blogs	EP, RA	S
'07	Devitt and Ahmad	C	N	Semantic	2	News	News	L
'07	Mei et al	C	Y	HMM	2	Blogs	N/A	L
'07	Ku et al	C	N	Statistic	C	News	NTCIR	L
'07	Chen et al	A	N	DT, SVM	2	Reviews (B)	N/A	S
'08	Annett and Kondrak	C	N	SVM, NB, ADT	2	Reviews (M)	IMDB	M
'08	He et al	C	N	Statistic (IR)	C	Blogs	TREC	M

Continued...

Year	Authors	Type	Topic	Algorithms	Range	Scope	Data	Scale
’08	Bestgen	C	N	Statistics (SO-LSA)	2	Words	N/A	L
’08	Fahrni and Klenner	C	Y	Statistic	C	Reviews (S)	N/A	L
’08	Shimada and Endo	C	Y	SVM OVA, ME, SVR	3, 6	Reviews (P)	N/A	L
’09	Zhang et al	A	Y	Corpus	C	News	N/A	L
’09	Miao et al	A	Y	Dictionary	2	Reviews (P)	N/A	M
’09	Zhu et al	A	Y	Dictionary	3	Reviews (S)	N/A	M
’09	Nadeau et al	C	N	LR, NB + Dictionary	4	Dreams	N/A	S
’09	Bodendorf	C	N	SVM OVA	3	Blogs	N/A	M
’09	Choi et al	C	Y	Clustering +Dictionary	3	News	NTCIR	S
’09	Lin and He	C	Y	LDA + Dictionary	2	Texts	IMDB	S
’09	Nowson	C	Y	SVM	2	Reviews (P)	N/A	S
’09	Melville	C	N	NB + Dictionary	2	Blogs	N/A	L
’09	Thet et al	C	Y	Dictionary	C	Reviews (M)	IMDB	L
’09	Prabowo and Thelwall	C	N	RuleBased, Dictionary, Statistic, SVM	2	Reviews(M,P)	IMDB,N/A	L
’09	Feng et al	A	Y	Dictionary	2	Blogs	N/A	L
’09	Lerman et al	A	N	Semantic	C	Reviews(P)	N/A	L
’09	Hare et al	C	Y	MNB, SVM	2, 3	Blogs	N/A	L
’09	Dasgupta and Ng	C	N	SVM + Clustering	2	Texts	IMDB,AZ	S
’09	Missen and Boughanem	C	Y	Semantic	C	Blogs	TREC	M
’09	Read and Carroll	C	N	Statistic	2	News, Reviews(M)	IMDB, GI	L, S
’09	Go et al	C	N	NB, ME, SVM	2	Microblogs	Twitter	L
’10	Bollen et al	A	N	OpinionFinder, Statistic (PMI)	C	Microblogs	Twitter	L
’10	Tumasjan et al	A	Y	Dictionary(LIWC)	C	Microblogs	Twitter	L
’10	Bifet and Frank	C	N	MNB, SGD, Hoeffding tree	2	Microblogs	Twitter	L
’10	Pak and Paroubek	C	N	MNB	3	Microblogs	Twitter	L

Table 1: An overview of the most popular sentiment extraction algorithms, used in Subjectivity Analysis.

6.2 Comparison of Methods

As can be seen in Figure 6, dictionary and machine learning approaches attract most of the attention in the research community. They have been evolving in parallel since the beginning of this decade, and it comes as no surprise that studies have started to compare their performance on different datasets. Below we present the most interesting comparisons and briefly discuss their results. A complete list of performance evaluations is reported in Table 2.

Chaovalit et al. (2005) performed an evaluation between the N-gram classifier and statistical approach methods on a dataset of movie reviews. In particular, their study showed the machine learning precision ranging from 66% (on the unseen data) to 85% (with 3-fold cross-validation), making it comparable to the 77% precision achieved with the unsupervised dictionary method.

Gindl et al. (2008) compared the precision between various dictionary and machine learning methods on web datasets (Amazon, IMDb, and TripAdvisor). The results demonstrated the superiority of the machine learning methods over the dictionary methods on all three datasets. The best results were achieved by the ME method, whose precision was in almost every case greater than 80%.

Another comparison between the most popular types of algorithms for sentiment extraction was made by Annett and Kondrak (2008), demonstrating that some semantics-based algorithms are able to keep up with machine learning methods in terms of precision, even though they do not require a computationally-demanding learning phase. In particular, a lexical algorithm utilizing WordNet polarity scores achieved a precision close to that of decision trees (60.4% versus 67.4%). Nevertheless, these algorithms do not substitute, but rather complement each other.

As was demonstrated by Prabowo and Thelwall (2009), only a combination of different kinds of classifiers is able to achieve a solid performance. In order to build their hybrid approach, they combined several rule-based classifiers with a statistical approach method and an SVM classifier. Doing so, they achieved a performance ranging from 83% to 91%, depending on the dataset.

We also point the interested reader to other studies that compare the performance of various Sentiment Analysis algorithms on different datasets (Prabowo and Thelwall, 2009; Chaovalit and Zhou, 2005; Annett and Kondrak, 2008).

However, a systematic comparative study that implements and evaluates all relevant algorithms under the same framework is still missing. Note that the performance results reported in Table 2 are not directly comparable to each other, because the evaluation framework and testing methodologies are not the same across the corresponding studies.

Paper	**Dataset**	**Sentiment Algorithm (Precision, %)**
Dave et al (2003)	AZ, CN	SVM (85.8 - 87.2) NB (81.9 - 87.0)
Hu and Liu (2004a)	AZ, CN	Semantic (84.0)
Turney (2002)	EP	Statistics (74.4)
Taboada et al (2006)	EP	PMI (56.8)
Turney and Littman (2003)	HM	SO-LSA (67.7 - 88.9) PMI (61.8 - 71.0)
	GI	SO-LSA (65.3 - 82.0) PMI (61.3 - 68.7)
Kamps et al (2004)	GI	Semantic (76.7)
Read and Carroll (2009)	GI	PMI (71.7) Semantic Space (83.8) Similarity (67.6)
	SemEval*	PMI (46.4) Semantic Space (44.4) Similarity (53.1)
	IMDB	PMI (68.7) Semantic Space (66.7) Similarity (60.8)
Gindl and Liegl (2008), average	AZ (N/A)	Dictionary (59.5 - 62.4) NB (66.0) ME (83.8)
	TA (N/A)	Dictionary (70.9 - 76.4) NB (72.4) ME (78.9)
	IMDB	Dictionary (61.8 - 64.9) NB (58.5) ME (82.3)
Pang et al (2002)	IMDB	NB (81.5) ME (81.0) SVM (82.9)
Chaovalit and Zhou (2005)	IMDB	N-Gram (66.0 - 85.0) PMI (77.0)
Goldberg and Zhu (2006)	IMDB	SVR (50.0 - 59.2) Graph (36.6 - 54.6)
Annett and Kondrak (2008)	IMDB	NB (77.5) SVM (77.4) ADTree (69.3)
Thet et al (2009)	IMDB	Dictionary (81.0)
Ku et al (2007)	NTCIR	Statistics (66.4)
Choi et al (2009)	NTCIR	Dictionary + Clustering ($\sim$70.0)
Osherenko and André (2007)	SAL*	SVM + Dictionary (34.5)
Yu and Hatzivassiloglou (2003)	TREC	Statistics (68.0 - 90.0)

Continued. . .

Paper	Dataset	Sentiment Algorithm (Precision, %)
Ku et al (2005)	TREC	Dictionary (62.0)
Missen and Boughanem (2009)	TREC	Semantic (MAP 28.0, P@10 64.0)
Yi et al (2003)	N/A	Dictionary (87.0 Reviews, 91.0 - 93.0 News)
Gamon (2004)	N/A	SVM (69.0 nearest classes, 85.0 farthest classes)
Kim and Hovy (2004)	N/A	Semantic (67.0 - 81.0)
Thomas et al (2006)	N/A	Multiple SVM (71.0)
Nadeau et al (2006)	N/A*	LR (35.0 - 50.0) NB + Dictionary (38.0)
Chen et al (2006)	N/A	DT (71.7) SVM (84.6) NB (77.5)
Devitt and Ahmad (2007)	N/A	Semantic (50.0 - 58.0, f-measure)
Shimada and Endo (2008)	N/A*	SVM OVA (58.4) ME (57.1) SVR (57.4) SIM (55.7)
Hare et al (2009)	N/A	MNB (75.1) SVM (74.4)
Zhu et al (2009)	N/A	Dictionary (69.0)
Bodendorf (2009)	N/A	SVM OVA (69.0)
Melville (2009)	N/A	NB + Dictionary (63.0 - 91.0)
Prabowo and Thelwall (2009)	N/A	SVM-only (87.3) SVM + RuleBased + Dictionary + Statistics (91.0)
Feng et al (2009)	N/A	Dictionary (65.0)
Go et al (2009)	TS	NB (82.7) ME (83.0) SVM (82.2)
Bifet and Frank (2010)	TS	MNB (82.5) SGD (78.6), Hoeffding tree (69.4)
	N/A	MNB (86.1) SGD (86.3) Hoeffding tree (84.8)
Pak and Paroubek (2010)	N/A	MNB (70.0) at recall value 60.0

Table 2: Precision of sentiment extraction for different implementations according to the data reported by authors. Due to the limited space, in this table we only report best-run results for the available datasets (which are also listed in Table 3). "N/A" means that the dataset is not publicly available. 3(5)-classes accuracy marked with *.

6.3 Specifics of Web Mining

The evaluations found in Yi et al (2003); Ku et al (2007); Dave et al (2003); Annett and Kondrak (2008) demonstrate that opinion data obtained from the web, are represented primarily in discrete or categorical form. This happens not only because of the use of machine learning tools, but also because ratings and opinion labels are represented by a limited number of categories on the web. Such availability of categorical training data favors the use of machine learning for such tasks as rating inference or review mining, and made machine learning tools the default choice for solving the Opinion Mining problem.

A side effect of the domination of these tools is that the sentiment classification task is mostly considered as a binary- or three-class classification problem, distinguishing among *positive*, *negative*, or *neutral* texts. However, it is not clear whether this approach is the winner. On the contrary, recent studies demonstrate the benefits of employing more complex (detailed) sentiment classifications (Tsytsarau et al, 2011; Thet et al, 2009). Moreover, it is not always possible to use supervised machine learning methods. For example, when there are no annotated training data (like in blog opinion retrieval), dictionary approaches that provide sentiment values on a continuous scale, become an interesting alternative.

Most of the works in Subjectivity Analysis assume a set of predefined topics when determining sentiments. These topics are specified either by keywords, or by restricting the collection of documents to only those that mention the chosen topics. In other words, the algorithms operate on the implicit assumption of a *single document - single topic* context. This situation changes when it is necessary to analyze sentiments expressed in free-form texts (e.g., weblogs), which may involve several topics. To solve this new problem, *single document - several topics* context, these methods should be extended with topic identification algorithms. Stoyanov and Cardie (2008) present an approach for opinion topic extraction that relies on the identification of topic-coreferent opinions. Alternatively, Mei et al. (2007) and Lin et al. (2009) propose to include sentiment variables into a probabilistic topic inference model.

6.4 Open Problems

The mining and analysis of opinions is a challenging and interdisciplinary task, which requires researchers from different domains to consolidate their efforts. A typical solution in this area requires fast and scalable information retrieval, text preprocessing and topic assignment, in order to run machine learning algorithms supported by the possible use of NLP tools.

We observe that both the performance and resolution of the Subjectivity Analysis algorithms have increased over time. The first algorithms that were proposed in the literature were effective at discriminating between two or among three classes of sentiments. As we mention in Section 3.3, switching to several opinion classes required a redesign of the employed machine learning methods (Pang and Lee, 2005), while continuous sentiment values are only obtainable by using dictionary-based methods. Based on this, we foresee that the increasing demand for the quality of sentiments will require the development of new methods that will inherit strong features from both the machine learning and the dictionary-based methods.

As we are interested in integrating recent developments of Opinion Mining, we need to develop a universal scale to represent opinions. For the sentiment analysis problem, the choice of the continuous scale in the range of [-1;1] seems to be a natural one, as it easily accommodates the discrete opinion categories (-1,0,1), and at the same time provides flexible opportunities for various mappings from the rating scale (e.g., rating stars). However, for conflicting opinions there is no such obvious choice. We need to represent differences in opinions that can not be directly mapped to real values. For example, the pair “the cat is *black* - it is a *white* cat” that features an obvious contradiction, can not be represented using +/-1, as the set containing just two colors (*black*, *white*) is not complete - there might also be *gray, red* and others.

Our study also reveals the need to address the problems of aggregating, managing, and analyzing sentiments in a large scale, and in an *ad hoc* fashion, much like the analysis opportunities offered by On-Line Analytical Processing (OLAP) in traditional data management. Such methods would only be possible if we will manage to solve sentiment aggregation problems with high efficiency. The latter would also depend on the successful introduction of a common rating scale. In order to make significant advances along the above directions, we need to introduce an appropriate framework, and formally define the corresponding problems.

Moreover, there is a need for novel techniques that will summarize and analyze the relevant information in a principled and systematic way. We anticipate the introduction of a collaborative framework that will further advance the state of the art and establish new targets for the next decade. Contradiction Analysis can possibly be the most demanding field for such a framework, as it utilizes most of the opinion mining methods, and at the same time defines its problems on data of various types, ranging from opposite sentiments to conflicting facts. We believe that it encompasses most of the challenges relevant to Subjectivity Analysis, and can be used as a reference target for the development of the framework mentioned above.

Finally, we note the lack of benchmarks in this area, which would greatly help its further development. Even though some datasets annotated with sentiments are available, they do not have the required precision and resolution. The problem is even more exacerbated when dealing with the most recent algorithms and applications, such as those relevant to Contradiction Analysis. In Table 3, we list the various datasets that have been used for Subjectivity Analysis (mainly Opinion Mining).

Regarding the contradictions between natural-language texts, the research in this direction is supported by the RTE challenge[5], which initiated a three-way classification task in 2008. In addition to the two-way classification between entailment and non-entailment, this task includes detection of contradiction as a part of non-entailment classification.

7 Conclusions

During the past decade, we have witnessed an increasing interest in the processing and analysis of unstructured data, with a special focus on Web text data. The wealth of information on the Web makes this endeavor not only rewarding in terms of newly produced knowledge, but also necessary, in order to exploit all this available information. We believe that the interest in mining Web data would only continue to grow, as new sources of such data emerge and attract more attention from users and researchers alike.

In this work, we presented an overview of a special class of web mining algorithms, that of Subjectivity Analysis. This is an area that started developing in the last years, and attracted lots of attention, because of its practical applications and the promise to uncover useful and actionable patterns from unstructured web data.

More specifically, we reviewed the most prominent approaches for the problems of *Opinion Mining* and *Opinion Aggregation*, as well as the recently introduced *Contradiction Analysis*. These have emerged as important areas of web data mining, and the trends of the past years show an increasing involvement of the research community, along with a drive towards more sophisticated and powerful algorithms. Our survey reveals these trends, identifies several interesting open problems, and indicates promising directions for future research.

Acknowledgments We would like to thank the anonymous reviewers and the editor for their numerous valuable comments, which helped to significantly improve the quality of the content and presentation of this survey.

[5] http://www.nist.gov/tac/2010/RTE/index.html

Name (and URL)	Year	Type	Pos	Neg	Neu	Range
GI - General Inquirer content analysis system *www.wjh.harvard.edu/~inquirer/*	2002	Words	N/A	N/A	N/A	D
IMDB - Movie Review Data v2.0 *www.cs.cornell.edu/People/pabo/movie-review-data/*	2004	Reviews	1,000	1,000	0	B
TREC - Blog Track *http://ir.dcs.gla.ac.uk/test_collections/blog06info.html*	2006	Blogs	3,215,171 total			N/A
AZ - Amazon Reviews *www.cs.uic.edu/~liub/FBS/sentiment-analysis.html*	2007	Reviews	4,554K	759K	525K	D
SemEval-2007 Affective Text Task *www.cse.unt.edu/~rada/affectivetext/*	2007	News	561	674	15	D
NTCIR-MOAT - *http://research.nii.ac.jp/ntcir/*	2008	News	26K total			D
SAL - Sensitive Artificial Listener corpus *www.vf.utwente.nl/~hofs/sal/*	2008	Speech	672 turns total			D
Irish Economic Sentiment Dataset *www.mlg.ucd.ie/sentiment/*	2009	News	2,608	3,915	3,080	D
Multi-Domain Sentiment Dataset v2.0 *www.cs.jhu.edu/~mdredze/datasets/sentiment/*	2009	Reviews	10,771	10,898	0	D
TS - Twitter Sentiment *http://twittersentiment.appspot.com/*	2009	Micro-blogs	800K	800K	0	B
TA - TripAdvisor.com *http://times.cs.uiuc.edu/~wang296/Data/*	2010	Reviews	183K	37K	26K	D
EP - Epinions - *www.epinions.com*	2010	Reviews	N/A	N/A	N/A	D
CN - C|net - *www.cnet.com*	2010	Reviews	N/A	N/A	N/A	D
RA - RateItAll - *www.rateitall.com*	2010	Reviews	N/A	N/A	N/A	D
ZD - ZDnet - *www.zdnet.com*	2010	Reviews	N/A	N/A	N/A	D

Table 3: An overview of the most popular opinion mining datasets and data sources. Under the columns “Pos”, “Neg” and “Neu” we list the approximate numbers of positive, negative and neutral labels respectively. The range of these labels is either binary, marked as “B”, or discrete, marked as “D”. “N/A” means that the dataset is not publicly available.

References

Annett M, Kondrak G (2008) A comparison of sentiment analysis techniques: Polarizing movie blogs. In: Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence, Canadian AI ’08, pp 25–35

Antweiler W, Frank MZ (2004) Is all that talk just noise? the information content of internet stock message boards. Journal of Finance 59(3):1259–1294

Archak N, Ghose A, Ipeirotis PG (2007) Show me the money!: deriving the pricing power of product features by mining consumer reviews. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD ’07, pp 56–65, DOI http://doi.acm.org/10.1145/1281192.1281202

Bermingham A, Smeaton AF (2010) Classifying sentiment in microblogs: is brevity an advantage? In: Huang J, Koudas N, Jones G, Wu X, Collins-Thompson K, An A (eds) CIKM, ACM, pp 1833–1836

Bifet A, Frank E (2010) Sentiment knowledge discovery in Twitter streaming data. In: Proceedings of the 13th International Conference on Discovery Science, Springer, Canberra, Australia, pp 1–15

Bollen J, Mao H, Zeng XJ (2010) Twitter mood predicts the stock market. CoRR abs/1010.3003

Carenini G, Ng RT, Zwart E (2005) Extracting knowledge from evaluative text. In: Proceedings of the 3rd international conference on Knowledge capture, ACM, New York, NY, USA, K-CAP '05, pp 11–18, DOI http://doi.acm.org/10.1145/1088622.1088626

Carenini G, Ng R, Pauls A (2006) Multi-document summarization of evaluative text. In: Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics, pp 3–7

Chaovalit P, Zhou L (2005) Movie review mining: a comparison between supervised and unsupervised classification approaches. Hawaii International Conference on System Sciences 4:112c, DOI http://doi.ieeecomputersociety.org/10.1109/HICSS.2005.445

Chen C, Ibekwe-SanJuan F, SanJuan E, Weaver C (2006) Visual analysis of conflicting opinions. In: IEEE Symposium on Visual Analytics Science and Technology, pp 59–66

Chen F, Tan PN, Jain AK (2009) A co-classification framework for detecting web spam and spammers in social media web sites. In: Proceeding of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM '09, pp 1807–1810, DOI http://doi.acm.org/10.1145/1645953.1646235

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research 43(3):345–354

Choudhury MD, Sundaram H, John A, Seligmann DD (2008) Multi-scale characterization of social network dynamics in the blogosphere. In: CIKM, pp 1515–1516

Church KW, Hanks P (1989) Word association norms, mutual information, and lexicography. In: Proceedings of the 27th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 76–83, DOI http://dx.doi.org/10.3115/981623.981633

Dave K, Lawrence S, Pennock D (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on World Wide Web, ACM, New York, NY, USA, WWW '03, pp 519–528, DOI http://doi.acm.org/10.1145/775152.775226

Devitt A, Ahmad K (2007) Sentiment polarity identification in financial news: A cohesion-based approach. In: 45th Annual Meeting of the Association of Computational Linguistics

Ekman P, Friesen WV, Ellsworth P (1982) What emotion categories or dimensions can observers judge from facial behavior? In: Emotion in the human face, Cambridge University Press, New York, pp 39–55

Ennals R, Byler D, Agosta JM, Rosario B (2010a) What is disputed on the web? In: Proceedings of the 4th ACM Workshop on Information Credibility on the Web, WICOW 2010, Raleigh, USA, April 27, 2010

Ennals R, Trushkowsky B, Agosta JM (2010b) Highlighting disputed claims on the web. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, USA, April 26-30, 2010

Esuli A, Sebastiani F (2006) Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC '06

Fahrni A, Klenner M (2008) Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In: Proceedings of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention, pp 60 – 63

Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA

Feng S, Wang D, Yu G, Yang C, Yang N (2009) Sentiment clustering: A novel method to explore in the blogosphere. In: Proceedings of the Joint International Conferences on Advances in Data and Web Management, Springer-Verlag, Berlin, Heidelberg, APWeb/WAIM '09, pp 332–344, DOI http://dx.doi.org/10.1007/978-3-642-00672-2_30

Galley M, McKeown K, Hirschberg J, Shriberg E (2004) Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, ACL '04, pp 669–676, DOI http://dx.doi.org/10.3115/1218955.1219040

Giampiccolo D, Dang HT, Magnini B, Dagan I, Cabrio E, Dolan B (2008) The fourth pascal recognizing textual entailment challenge. In: Proceedings of the First Text Analysis Conference, TAC '08

Gindl S, Liegl J (2008) Evaluation of different sentiment detection methods for polarity classification on web-based reviews. In: Proceedings of the 18th European Conference on Artificial Intelligence, pp 35–43

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Tech. rep., Stanford University

Godbole N, Srinivasaiah M, Skiena S (2007) Large-scale sentiment analysis for news and blogs. In: Proceedings of the International Conference on Weblogs and Social Media, ICWSM '07

Goldberg A, Zhu X (2006) Seeing stars when there aren't many stars: Graph-based semisupervised learning for sentiment categorization. In: TextGraphs Workshop On Graph Based Methods For Natural Language Processing

Harabagiu S, Hickl A, Lacatusu F (2006) Negation, contrast and contradiction in text processing. In: AAAI'06: Proceedings of the 21st national conference on Artificial intelligence, pp 755–762

He B, Macdonald C, He J, Ounis I (2008) An effective statistical approach to blog post opinion retrieval. In: CIKM, pp 1063–1072

Hillard D, Ostendorf M, Shriberg E (2003) Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In: HLT-NAACL

Hoffman T (2008) Online reputation management is hot - but is it ethical? Computerworld

Horrigan JA (2008) Online shopping. Pew Internet and American Life Project Report

Hu M, Liu B (2004a) Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '04, pp 168–177, DOI http://doi.acm.org/10.1145/1014052.1014073

Hu M, Liu B (2004b) Mining opinion features in customer reviews. In: Mcguinness DL, Ferguson G, Mcguinness DL, Ferguson G (eds) AAAI, AAAI Press / The MIT Press, pp 755–760

Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the international conference on Web search and web data mining, ACM, New York, NY, USA, WSDM '08, pp 219–230, DOI http://doi.acm.org/10.1145/1341531.1341560

Kamps J, Marx M, Mokken RJ, Rijke MD (2004) Using wordnet to measure semantic orientation of adjectives. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC '04, vol IV, pp 1115–1118

Kim HD, Zhai C (2009) Generating comparative summaries of contradictory opinions in text. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM '09, pp 385–394, DOI http://doi.acm.org/10.1145/1645953.1646004

Kim SM, Hovy E (2004) Determining the sentiment of opinions. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, COLING '04, p 1367, DOI http://dx.doi.org/10.3115/1220355.1220555

Koppel M, Schler J (2006) The importance of neutral examples for learning sentiment. Computational Intelligence 22(2):100–109

Ku LW, Liang YT, Chen HH (2006) Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs

Ku LW, Lo YS, Chen HH (2007) Using polarity scores of words for sentence-level opinion extraction. In: Proceedings of NTCIR-6 Workshop Meeting, pp 316–322

Leung CWK, Chan SCF, Chung FL (2006) Integrating collaborative filtering and sentiment analysis: A rating inference approach. In: ECAI 2006 Workshop on Recommender Systems, pp 62–66

Lim E, Liu B, Jindal N, Nguyen V, Lauw W (2010) Detecting product review spammers using rating behaviors. In: CIKM, Toronto, ON, Canada

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceeding of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM '09, pp 375–384, DOI http://doi.acm.org/10.1145/1645953.1646003

Liu B (2010) Sentiment analysis and subjectivity. In: Indurkhya N, Damerau FJ (eds) Handbook of Natural Language Processing, Second Edition, CRC Press, Taylor and Francis Group, Boca Raton, FL, iSBN 978-1420085921

Liu B, Hu M, Cheng J (2005) Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the 14th international conference on World Wide Web, ACM, New York, NY, USA, WWW '05, pp 342–351, DOI http://doi.acm.org/10.1145/1060745.1060797

Liu H, Lieberman H, Selker T (2003) A model of textual affect sensing using real-world knowledge. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, IUI '03, pp 125–33

Liu J, Birnbaum L, Pardo B (2009) Spectrum: Retrieving different points of view from the blogosphere. In: Proceedings of the Third International Conference on Weblogs and Social Media

Lu Y, Tsaparas P, Ntoulas A, Polanyi L (2010) Exploiting social context for review quality prediction. In: Proceedings of the 19th international conference on World wide web, ACM, New York, NY, USA, WWW '10, pp 691–700, DOI http://doi.acm.org/10.1145/1772690.1772761

de Marneffe MC, Rafferty AN, Manning CD (2008) Finding contradictions in text. In: Proceedings of ACL: HLT, Association for Computational Linguistics, Columbus, Ohio, ACL '08, pp 1039–1047

McArthur R (2008) Uncovering deep user context from blogs. In: AND, pp 47–54

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

Miao Q, Li Q, Dai R (2009) Amazing: A sentiment mining and retrieval system. Expert Syst Appl 36(3):7192–7198, DOI http://dx.doi.org/10.1016/j.eswa.2008.09.035

Missen MM, Boughanem M (2009) Using wordnet's semantic relations for opinion detection in blogs. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR '09, pp 729–733, DOI http://dx.doi.org/10.1007/978-3-642-00958-7_75

Morinaga S, Yamanishi K, Tateishi K, Fukushima T (2002) Mining product reputations on the web. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, KDD '02, pp 341–349, DOI http://doi.acm.org/10.1145/775047.775098

Mullen T, Malouf R (2006) A preliminary investigation into sentiment analysis of informal political discourse. In: AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs

Nadeau D, Sabourin C, de Koninck J, Matwin S, Turney P (2006) Automatic dream sentiment analysis. In: Proceedings of the Workshop on Computational Aesthetics at the 21st National Conference on Artificial Intelligence, AAAI-06

Osherenko A, André E (2007) Lexical affect sensing: Are affect dictionaries necessary to analyze affect? In: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction, Springer-Verlag, Berlin, Heidelberg, ACII '07, pp 230–241, DOI http://dx.doi.org/10.1007/978-3-540-74889-2_21

Pado S, de Marneffe MC, MacCartney B, Rafferty AN, Yeh E, Manning CD (2008) Deciding entailment and contradiction with stochastic and edit distance-based alignment. In: Proceedings of the First Text Analysis Conference, TAC '08

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) LREC, European Language Resources Association

Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp 271–278

Pang B, Lee L (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: ACL

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

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: EMNLP 2002, pp 79–86

Prabowo R, Thelwall M (2009) Sentiment analysis: A combined approach. Journal of Informetrics 3(2):143–157, DOI 10.1016/j.joi.2009.01.003

Read J, Carroll J (2009) Weakly supervised techniques for domain-independent sentiment classification. In: Proceeding of the International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, ACM, New York, NY, USA, TSA '09, pp 45–52, DOI http://doi.acm.org/10.1145/1651461.1651470

Riloff E, Wiebe J, Phillips W (2005) Exploiting subjectivity classification to improve information extraction. In: Veloso MM, Kambhampati S (eds) AAAI, AAAI Press / The MIT Press, pp 1106–1111

Shimada K, Endo T (2008) Seeing several stars: A rating inference task for a document containing several evaluation criteria. In: PAKDD, pp 1006–1014

Stoyanov V, Cardie C (2008) Topic identification for fine-grained opinion analysis. In: Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, Coling '08, pp 817–824

Taboada M, Anthony C, Voll K (2006) Methods for creating semantic orientation dictionaries. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC '06, pp 427–432

Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. Expert Syst Appl 36(7):10,760–10,773, DOI http://dx.doi.org/10.1016/j.eswa.2009.02.063

Thet TT, Na JC, Khoo CS, Shakthikumar S (2009) Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In: Proceeding of the International

CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, TSA '09

Thomas M, Pang B, Lee L (2006) Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: EMNLP, pp 327–335

Tsytsarau M, Palpanas T, Denecke K (2010) Scalable discovery of contradictions on the web. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, USA, April 26-30, 2010

Tsytsarau M, Palpanas T, Denecke K (2011) Scalable detection of sentiment-based contradictions. In: First International Workshop on Knowledge Diversity on the Web, Colocated with WWW 2011, Hyderabad, India, March 28-31, 2011

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Cohen WW, Gosling S (eds) ICWSM, The AAAI Press

Turney P, Littman M (2003) Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21:315–346

Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on ACL, Association for Computational Linguistics, Morristown, NJ, USA, ACL '02, pp 417–424, DOI http://dx.doi.org/10.3115/1073083.1073153

Varlamis I, Vassalos V, Palaios A (2008) Monitoring the evolution of interests in the blogosphere. In: ICDE Workshops, IEEE Computer Society, pp 513–518

Voorhees EM (2008) Contradictions and justifications: Extensions to the textual entailment task. In: Proceedings of ACL: HLT, Association for Computational Linguistics, Columbus, Ohio, ACL '08, pp 63–71

Wiebe J, Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: CICLing-2005

Wiebe J, Wilson T, Bell M (2001) Identifying collocations for recognizing opinions. In: Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation, Association for Computational Linguistics, ACL '01, pp 24–31

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT/EMNLP, The Association for Computational Linguistics

Yi J, Nasukawa T, Bunescu R, Niblack W (2003) Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of the IEEE International Conference on Data Mining, ICDM '03

Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Collins M, Steedman M (eds) EMNLP, Sapporo, JP, EMNLP '03, pp 129–136, DOI http://portal.acm.org/citation.cfm?id=1119355.1119372

Zhang J, Kawai Y, Kumamoto T, Tanaka K (2009) A novel visualization method for distinction of web news sentiment. In: Vossen G, Long DDE, Yu JX (eds) WISE, Springer, Lecture Notes in Computer Science, vol 5802, pp 181–194

Zhang W, Yu CT, Meng W (2007) Opinion retrieval from blogs. In: Silva MJ, Laender AHF, Baeza-Yates RA, McGuinness DL, Olstad B, Olsen ØH, Falcão AO (eds) CIKM, ACM, pp 831–840

Zhou L, Chaovalit P (2008) Ontology-supported polarity mining. J Am Soc Inf Sci Technol 59:98–110, DOI 10.1002/asi.v59:1

Zhu J, Zhu M, Wang H, Tsou BK (2009) Aspect-based sentence segmentation for sentiment summarization. In: Proceeding of the International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, ACM, New York, NY, USA, TSA '09, pp 65–72, DOI http://doi.acm.org/10.1145/1651461.1651474