





## **Deliverable D7.1:**

# **State of the Art on Opinion Mining**

## **UbiPOL**



**Project reference: INFSO-ICT-248010**

	<b>UBIPOL – STATE OF THE ART ON OPINION MINING</b> Del. no: D7.1	
<b>INFSO-ICT-248010</b>	<b>Version: Approved</b>	<b>20/01/2011</b>

**Deliverable type:** R

**Classification:** RE

**Work package and task:** WP7, T7.1

**Responsibility:** SU

**Executive summary:** This document presents the state of the art on opinion mining which will be the basis of our research for the UbiPOL opinion mining engine. Opinion mining is the process of automatically extracting what the writer thinks about a specific topic. In the context of UbiPOL, we are going to target the comments entered by the citizens about policy issues and try to extract the opinion of citizens about these policy issues as a feedback mechanism for the policy making process. We plan to exploit domain ontologies to incorporate semantics into the opinion mining process. In this report, we provide a survey of existing approaches for opinion mining including polarity analysis together with evaluation metrics. We also surveyed the state of the art natural language processing tools that can be used for opinion mining.

## Amendment History

date	issue	status	Author
20-11-2010	Revision Level 0.A	The first draft of the document by SU.	Dilek Tapucu, Max Zimmerman, Zaigham Faraz Siddiqui
20-12-2010	Revision Level 0.B	Introduction was updated by SU.	Yucel Saygın
28-12-2010	Revision Level 1.A	The first completed draft for review by peers	Dilek Tapucu, Max Zimmerman, Zaigham Faraz Siddiqui
20-01-2011	Revision Level 1.B	Updated according to review comments.	Yucel Sayg�n

<b>1.</b>	<b>INTRODUCTION .....</b>	<b>5</b>
<b>2.</b>	<b>EXISTING TECHNIQUES FOR OPINION EXTRACTION.....</b>	<b>7</b>
<b>2.1.</b>	<b>Pure Text-Based Opinion Mining .....</b>	<b>7</b>
<b>2.2.</b>	<b>Ontology-Based Opinion Mining.....</b>	<b>7</b>
<b>3.</b>	<b>ANNOTATION METHODS .....</b>	<b>9</b>
<b>4.</b>	<b>FINDING POLARIZED WORDS .....</b>	<b>10</b>
<b>5.</b>	<b>LEARNING OPINIONS FROM TWEETS.....</b>	<b>11</b>
<b>6.</b>	<b>OPINION SUMMARISATION .....</b>	<b>13</b>
<b>6.1.</b>	<b>Sentiment Lexicon.....</b>	<b>13</b>
<b>6.2.</b>	<b>Extraction of Relevant Aspects .....</b>	<b>14</b>
<b>7.</b>	<b>PREDICTING SENTIMENT ORIENTATION .....</b>	<b>16</b>
<b>8.</b>	<b>UNDERLYING TOOLS .....</b>	<b>17</b>
<b>8.1.</b>	<b>Protégé.....</b>	<b>17</b>
<b>8.2.</b>	<b>OnTeA .....</b>	<b>17</b>
<b>8.3.</b>	<b>GATE .....</b>	<b>17</b>
<b>9.</b>	<b>EVALUATION METRICS FOR SENTIMENT DETECTION .....</b>	<b>18</b>
<b>10.</b>	<b>REFERENCES.....</b>	<b>19</b>

---

## 1. Introduction

---

Nowadays, on the Web people express their opinions and this new source of information contains valuable sources of information for marketing intelligence and many other applications. Sentiment analysis or opinion mining is a growing research area both in natural language processing, and information retrieval communities. Sentiment analysts are turning their eyes on the web in order to help organizations and individuals to gain such information effectively and easily. Thus, new techniques are now being developed to exploit these sources.

A very broad overview of the existing work was presented in the literature. Researchers focus on extracting the affective content of a textual document from the detection of expressions of “bag of sentiment words” at different levels of granularity. The challenge here is to correctly classify a document's viewpoint (or polarity) as positive, negative or somewhere in-between. (Pang and Lee, 2008) described existing techniques and approaches for an opinion-oriented information extraction. (Turney and Littman, 2003) worked on the classification of documents into positive or negative categories, (Wiebe et al., 2004) classified text into subjective or objective, (Esuli and Sebastiani 2006), showed opinion mining consists both in searching for the opinions or sentiments expressed in a document and in acquiring new methods to automatically perform such analysis and (Wloka et al. 2007) analyzed and classified emotions like joy, anger or grief.

In this context our goal is to show the role of domain ontology for extracting features. Our idea is to produce polarity ontology to show user opinion. We want to identify opinion sentences in each review and decide whether each opinion sentence is positive or negative. (Note that these opinion sentences must contain one or more domain knowledge features.) With this view, we want to conduct experimental evaluations on a set of real posts. We noticed that similar approach is used in (Go et al., 2009) and in (Read, 2005). Authors used Twitter to collect training data and then to perform a sentiment search. They constructed corpora by using emoticons to obtain “positive” and “negative” samples, and then used various classifiers. Also in (Yang et al., 2007), the authors showed web-blogs to construct a corpora for sentiment analysis and used emotion icons assigned to blog posts as indicators of users' mood.

In this work to address the task of extracting opinions from a given document collection, Twitter sentences will use as real posts. Related reasons are given below:

- Twitter platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions.
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
- Text posts contain positive and negative emotions.

In this report our objective is to survey the various directions of research and tools for opinion mining, highlighting in particular the wide range of Information Extraction (IE) problems. We want to show that many IE solutions can significantly benefit from the wealth of work on managing structured data in the semantic web community. Finally, we hope that examining all these researches can in turn help us gain valuable insights into managing ontological data in this Internet-centric world.

This report was organized in the following sections: paper, we first make a survey of the various models then discuss the related work on opinion mining. Later we draw a picture of their relative positions with our proposed work and describe the steps of our work. Finally, we give discussion with the directions for future work.

## 2. Existing Techniques for Opinion Extraction

---

Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources [4]. There are two distinct approaches that are “Text based” and “Ontology based” presented in the literature to allow information extraction.

- “Text based information Extraction (TBIE) is a key NLP technology for automatically extracting specific types of information from text or populate knowledge bases”. Authors (Mooney and Bunescu 2005) give a good survey of current techniques.
- “Ontology-Based IE (OBIE) is the process of identifying relevant concepts, properties, and relations that come from in text or other sources expressed in ontology. “It is one of the technologies used for semantic annotation, which is essentially about assigning to entities in the text links to their semantic descriptions”.

There are two main difference between text based and ontology based IE; the important difference is the use of a formal ontology rather than a flat lexicon or gazetteer structure. This may also involve reasoning. The other one is OBIE not only finds the (most specific) type of the extracted entity, but it also identifies it, by linking it to its semantic description in the ontology.

Opinion Mining or Sentiment Analysis which is a recent discipline at the intersection of information extraction, computational linguistics, and text mining is concerned with identifying and classifying opinions in unstructured texts. It refers to the process to analyses and extracts occurrence of concepts and the sentiments associated with more refined information. Its aim is to extract relevant information about public opinion from unstructured data. There are two different kinds of approach consist of the literature that are described below.

### ***2.1. Pure Text-Based Opinion Mining***

“Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns [10]”. In order to properly capture opinion and sentiment expressed in a text, system needs a deep text processing approach.

Textual information in the world categorize into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. Opinions are usually subjective expressions that describe people’s sentiments, appraisals or feelings toward entities, events and their properties. The concept of opinion is very broad. Initial work (Hu and Liu, 2004; Jindal and Liu, 2006) on the product review data showed how opinions annotated directly. But they do not annotate comments that express opinions implicitly. Later, natural language processing techniques combined with machine learning algorithms allow building or extending ontologies in a semi-automatic manner. And background knowledge in form of ontologies enhances the performance of classical text mining tasks such as text classification and text clustering. Both the top-down and bottom-up approaches combine support for ontology lifecycle development.

### ***2.2. Ontology-Based Opinion Mining***

Tim Berners-Lee, the Web founder, puts forward the idea of the Semantic Web for the first time in 1998. Semantic Web is an expansion of the current Web rather than a new Web, it focus on how to make the web information understand and deal with by computer, that is, the web information with semantic elements. The use of ontologies has been shown to be particularly effective in specialized knowledge domains. Ontology population is a crucial part of knowledge base construction and maintenance that enables us to relate text to ontologies, providing on the one hand a customized ontology related to the data and domain with which we are concerned, and on the other hand a richer ontology which can be used for a variety of semantic web-related tasks such as knowledge management, information retrieval, question answering, semantic desktop applications, and so on. Ontology help to define concepts, relationships and entities that describe a domain with unlimited number of terms. This set of terms can be a significant and valuable lexical resource for extracting explicit and implicit features. For example:

If the concept customer is linked to the concept restaurant by the relation to eat in, a positive opinion towards the restaurant can be extracted from the review: we eat well. Similarly, if the concept restaurant is linked to the concept landscape with the relation to view, a positive opinion can be extracted towards the look out of the restaurant from the following review: very good restaurant where you can see the most beautiful peak of the Pyrenees.

The traditional way of the document indexing based on keywords can not work well in the network environment, which lacks the ability of semantic deducing. The semantic models provide users with a high level conceptualization of the information, while at the same time enabling them to focus on specific parts of the model. A more detailed description of the annotation process is given in Section 2.3.

Issues are also in free text format. They contain meta-information, such as time of publication and location. Issues can be categorized into a list of predefined topics. Issues and their relevant aspects are depicted in the UbiPOL ontology.



---

### 3. Annotation Methods

---

An annotation is a form of meta-data attached to a particular section of document content. In the literature, text has been annotated at a number of different levels of granularity including document, sentence, and phrase-level. For the creation of annotations manual, semiautomatic, and automatic approaches also exist.

The purpose of the semantic annotation process is to annotate words with domain-specific information. It helps to bridge the ambiguity of the natural language when expressing notions and their computational representation in a formal language. By telling a computer how data items are related and how these relations can be evaluated automatically, it becomes possible to process complex filter and search operations. Semantic Annotation approach goes one level deeper: It enriches the unstructured or semi-structured data with a context that is further linked to the structured knowledge of a domain, and, it allows results that are not explicitly related to the original search.

Ontology-based annotation prototypes support semantic annotating and adding semantic information to the web data, which make machines and humans understand. The use of human annotation is typically part of the process necessary in a supervised learning approach to sentiment analysis. The annotator generally annotates a document with the polarity of sentiment that contains with regard to a pre-defined topic. Semantic annotation technologies are the key in realizing the Semantic Web assuming, and directly determine the availability and scale of Semantic Web which is also one of the core issues in the Semantic Web research and application.

In the next section we concentrate on sentiment analysis tools to create features that enable us to carry out classification of short and long text reviews in any application domain. We believe that without appropriate tools, identifying and tracking public opinion and sentiments on particular topics is far from trivial.

---

## 4. Finding Polarized Words

---

The author of a text may express positive or negative sentiment. The smallest unit of text are the words contained in it. The polarity of words or phrases may express the writers sentiment on the matter he is discussing about. Adjectives and adverbs are the dominant type of words for sentiment words extraction and orientation identification in current research. Extracting those words is usually the first phase in finding the semantic orientation of a whole document. The main approaches to identify the semantic orientation of a sentiment word/phrases are statistical-based or lexicon-based.

A well-known statistical-based approach has been proposed by Hu and Liu [Hu and Liu, 2004]. They use the semantic orientation of synonyms and antonyms to predict the orientation of adjectives. Synonyms and antonyms for a adjective are found in WordNet. In WordNet, adjectives are set into relations of anonyms and synonyms, i.e. each adjective has two list, one contains anonyms and the other the synonyms according to adjective.

Rather than using adjectives Nasukawa and Yi [Nasukawa and Yi, 2003] consider verbs as sentiment expression for their sentiment analysis. They classify sentiment-related verbs into two types. The first type directs either positive or negative sentiment with respect to their arguments like "beat" in "X beats Y", where X and Y are the arguments. "beat" has then a positive sentiment when the phrase "X beats Y" is something positive according to the content of the text. For instance, "X beats Y" implies a positive sentiment when X and Y are soccer teams and the text is mainly written about team X. The second type transfers the sentiment of a verb among their arguments like "is" in "X is good", i.e. "is" is denoted as positive only in terms of the arguments "X" and "good".

Davidov and Rappoport [Davidov and Rappoport, 2006] uses the frequency of words in a set of documents to differ between words that express content, i.e. words that have a lot of meaning in terms of the documents, and words that have a high frequency but no meaning, i.e. words that occur more than TH times per million words, where TH is a threshold. Normally high frequency words are words like "of, or, and" etc. The two terms are then used to extract four types of patterns out of the words: CHC, CHCH, CHHC and HCHC, where C is a content word and H is a high frequency word. That is, a pattern contains two content. From the extracted patterns, the asymmetric patterns are filtered, i.e. only those patterns are kept which contain content words that appear at both positions of the pattern. For instance, "the house is small" (HCHC) and "a small white house" (HCHC) where the type of pattern is the same but the positions of "small" and "house" are different. Symmetric patterns are then used to find categories of words such as a adjective category that contains all words which express much happiness (amazing, fantastic, awesome, etc. ).

Recently Wu and Wen [Wu and Jin, 2010] proposed a method to disambiguate adjectives that are ambiguous such as large, small, high, low. Those adjectives express in some context a positive opinion, e.g. large income, and in a other context a negative opinion, e.g. large amount of taxes. Wu and Wen polarize nouns to find the orientation of adjectives as they assume that noun-adjective phrases (pressure is small) contain the orientation of a adjective.

## 5. Learning Opinions From Tweets

---

Twitter is a micro-blogging service built to discover what is happening at any moment in time, anywhere in the world. Tweets (messages in twitter) are different from reviews because of their purpose: while reviews represent summarized thoughts of authors, tweets are more casual and limited to 140 characters of text. Generally, tweets are not as thoughtfully composed as reviews. Tweets are a new source of information for opinion mining techniques. Learning opinions from tweets has become a challenging topic. The goal is to classify tweets into two categories depending on whether they convey positive or negative feeling.

Go et al. [Go, Bhayani, and Huang, 2009] first proposed a approach for automatically classifying the sentiment of Twitter messages. They uses emoticons to label a training data consists of tweets. Emoticons are visual cues that are associated with emotional states (e.g. some emoticons express happiness and some express sadness). This approach was introduced by Read [Read, 2005]. Due to the easy extraction of large amounts of tweets which contain emoticons, Go et al. [Go, Bhayani, and Huang, 2009] collected a huge training data set with labeled tweets on which they learned a naive Bayes classifier and a SVM. Words are represented as unigrams, i.e. each word is considered as a single word. The classifiers were run against a test set of tweets which may or may not have emoticons in them.

Bifet and Eibe [Bifet and Eibe, 2010] label tweets by regarding the authors provided sentiment indicators such as emoticons. Changing sentiment is then implicit in the use of various types of emoticons. Hence, Bifet and Eibe [Bifet and Eibe, 2010] use these to label a training data set. The labeled tweets are used to train a Naive Bayes classifier and a SVM. But in contrast to Go et al.

[Go, Bhayani, and Huang, 2009], Bifet and Eibe [Bifet and Eibe, 2010] came up with a data stream model that is able to deal with data whose nature or distribution changes over time. Pak and Paroubek [Pak and Paroubek, 2010a] use two twitter corpora to train a classifier that is able to classify documents into positive, negative and neutral classes. One twitter corpus contains a mixture of positive and negative annotated tweets (subjective texts) while the other is a set only with neutral tweets (objective texts). The positive and negative tweets are tweets which contain emoticons such as (":-)") or ":-(" ). Those emoticons have been used to annotate the tweets with a positive or a negative label. Moreover, they used part of speech tagging to tag the words. Based on the tagged words and the labeled training dat, Pak and Paroubek [Pak and Paroubek, 2010a] observed that objective texts tend to contain more common and proper nouns (NPS, NP, NNS), while authors of subjective texts use more often personal pronouns (PP,PP). They uses Naive Bayes and an SVM classifier to learn a model that distinguishes between subjective and objective texts.

Another work of Pak and Paroubek [Pak and Paroubek, 2010b] proposed using twitter to disambiguate sentiment ambiguous adjectives such as large, small, high, low. Those adjectives express in some context a positive opinion, e.g. large income, and in a other context a negative opinion, e.g. large amount of taxes. They train a classifier based on training data labeled by emoticons (similarly to Go et al. [Go, Bhayani, and Huang, 2009]) to classify texts into positive or negative sets. In contrast to [Go, Bhayani, and Huang, 2009], they use a combination of unigrams, bigrams and trigrams to model the words. The classes are then used to determine statistics that disambiguate sentiment ambiguous adjectives.

Davidov et al. [Davidov, Tsur, and Rappoport, 2010] use a twitter corpus to build a semi-supervised approach to recognize sarcastic sentences. Each tweet was annotated by hand from three different annotators. For annotating, they used a rating system consisting of 5 digits where the 1 and 2 represent no sarcasm and the rest represent an increasing amount of sarcasm. The obtained training set is then used to derive word frequencies (similarly to Davidov and Rappoport [Davidov and Rappoport, 2006]) which are used to rate tweets.

Bollen et al. [Bollen, Mao, and Zeng, 2010] investigated whether the mood of posted tweets is correlated to the value of Dow Jones Industrial Average (DJIA). They used an OpinionFinder [Wilson, Hoffmann, Somasundaran, Kessler, Wiebe, Choi, Cardie, Riloff, and Patwardhan, 2005] to identify subjective sentences and to mark various aspects of the subjectivity in these sentences including the source (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments; and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind and Happy). The output of OpinionFinder is a positive vs. negative time series of public mood, i.e. it comes up with a time series that depicts the daily mood by displaying the time at the x-axis and the mood at the y-axis. GPOMS determines six time series, one for each dimension different. A statistical hypothesis test is used to determine whether the seven time series are useful in predicting the changes in DJIA over time.

---

## 6. Opinion Summarisation

---

With the advancement and rapid expansion of Internet usage, more and more people are now using online forums, personal blogs, vendors' website and etc to discuss and express their opinions on topics ranging from everyday issues to what they buy online, e.g., products' review submitted by customers at online stores, movies' review by professional critics and general viewers, discussions on online forums ranging from fashion to governmental policies. Such data are usually un-structured and huge. Conventional methods used in information retrieval and text mining are mainly concerned with the overall information presented and have limited applicability in this domain. For example, a review about a laptop not only provides an overall score/sentiment but also provides separate sentiments on its individual aspects<sup>1</sup> such as battery life, mobility, processing power and etc. The individual sentiments hidden in review texts cannot be detected with methods that extract the overall text sentiment only. Thus, a more in-depth analysis that takes smaller textual fragments into consideration is imperative. This has led the researchers to combine methods from the domain of information retrieval and natural language processing to perform information extraction at higher granularity, e.g., paragraphs and sentences [Hu and Liu, 2004, Lerman et al., 2009, Thet et al., 2008].

The focus of this section is mainly on the work related to opinion summarisation, which is now a sub-field in information retrieval as well. Some earlier work in opinion mining deals with classifying the overall sentiment present in a review [Cui et al., 2006, Dave et al., 2003]. These works divide the data into train/test segments. Their main task in these methods is first to learn the polarity or sentiment orientation of each word, then apply the learned model on the test corpus.

One of the first works that identified the importance of a more granular view in opinion summary was the work of Hu and Liu [Hu and Liu, 2004]. They explained that in online reviews of products by customers (e.g., at amazon.com) the overall rating is of very little use. Quite often, reviewers tend to discuss individual aspects and features of a product and rate them independently. To capture the underlying opinion of the review, opinion about each individual aspect must be acquired independently. They can then be merged together into a meaningful summary in which each individual aspect is associated with how customers rate it individually. The framework usually employed for such aspect-based summarisation is:

- Identification of all the review fragments that express sentiment
- Identification of all the relevant aspects about which sentiment is expressed
- Use of a score function to compute the polarity of the sentiment expressed in the fragment.

Methods that comply with this framework include [Hu and Liu, 2004, Liu et al., 2005, Thet et al., 2008, Pang and Lee, 2008, Blair-Goldensohn et al., 2008, Lerman et al., 2009, Conrad et al., 2009, Potthast and Becker, 2010, Somprasertsri and Lalitrojwong, 2010, Li et al., 2010].

### 6.1. Sentiment Lexicon

Unlike the methods of Cui et al. and Dave et al. [Cui et al., 2006, Dave et al., 2003] that consider overall sentiment only, the methods that comply with the above defined framework do not require a training corpus to compute the polarity or inherent negativity or positivity of a term. Instead, they use a so-called sentiment lexicon. To build this sentiment lexicon, a small amount of seed words (containing words of negative, positive and neutral polarity) are

retrieved using the WordNet for synonyms and antonyms. Blair-Goldensohn et al. and Potthast and Becker [Blair-Goldensohn et al., 2008, Potthast and Becker, 2010] also employ methods that additionally weigh each word with a confidence measure that measures how likely is it for the word to have the designated polarity. More recently, Baccianella et al. [Baccianella et al., 2010] presented a lexical resource called SentiWordNet3.02 which annotates each synset of WordNet with positive ( $Pos(s)$ ), negative ( $Neg(s)$ ) and neutral ( $Obj(s)$ ) notions.

## 6.2. Extraction of Relevant Aspects

The main task of works is basically the extraction of the different aspects regarding a subject (e.g., product, policy, event and etc.) about which opinions are expressed and almost all the methods rely on part-of-speech (POS) tagging for the identification of the aspects. Method of Hu and Liu [Hu and Liu, 2004] performs mining of the opinions from product reviews. After the POS tagging, they use a classification rule miner (CBA) to identify different aspects of the products. They only consider noun phrases for this purpose and assume that users usually converge when talking about aspects. Any discovered itemset that is frequent (with a support of at least 1%) is treated as an aspect of the product. A pruning step is then performed to remove the infrequent features. The same authors proposed an improvement to their original method Liu et al. [Liu et al., 2005] by extracting aspects from pros and cons from shorter sentences.

Methods of Zhuang et al. and Thet et al. [Zhuang et al., 2006, Thet et al., 2008] specifically deals with the sentiment classification of movie reviews. Zhuang et al. [Zhuang et al., 2006] provides aspect classes, which are lists of keywords that may be potential features, to their method. For example, aspect class ST is related to screenplay, story, script and aspect class PAC is related to actors, actresses, supporting cast. They also crawl the imdb site to acquire a complete list of cast to ease the aspect extraction process for proper names of the actors. They then use a dependency grammar graph to identify explicit opinion-aspect pairs, e.g., "the movie is a masterpiece" or "visual affects are excellent" (see Table 1). For identifying implicit aspects, e.g., "I wanted it to end as soon as possible" (movie, boring), they use a simple hard coded technique that can identify few aspects only. Method of Thet et al. [Thet et al., 2008] uses a similar method which can perform pronoun resolutions as well. They also employ a rule-based approach that automatically tags the sentences into different aspects, e.g., overall, cast, storyline and etc.

**Table 1: Aspect-opinion pairs from Zhuang et al. [Zhuang et al., 2006] (aspects are bold faced while opinions are emphasized)**

<i>great</i> JJ				<b>cast</b> NN
	<i>amod</i>			
<b>script</b> NN				<i>fails</i> VB
	<i>nsubject</i>			
<b>movie</b> NN	<i>nsubject</i>	<i>is</i> VB	<i>dobj</i>	<i>masterpiece</i> NN

Blair-Goldensohn et al. [Blair-Goldensohn et al., 2008] split the aspect extraction into two steps: dynamic and static aspect extraction. The dynamic extraction is based on the method of Hu and Liu [Hu and Liu, 2004]. They also employ heuristic methods to aid in aspect extraction. For example, an adjective usually precedes noun phrase (usually an aspect), such as, "great picture quality". They also drop those aspects that do not have enough sentiment-bearing words close to them. In the static extraction phase, the method takes as input a list of

coarse-grained aspects that may be of potential interest. For example, in a dataset with reviews about hotels, such aspects can be "food", "decor", "service".

**Table 2: A sample questionnaire from Conrad et al. [Conrad et al., 2009]**

Questions
What feature do people like about Vista?
What features do people dislike about Vista?

The work of Conrad et al. [Conrad et al., 2009] specifically deals with the problem of summarising blog entries on subjects of law. Unlike other works, where aspects of a review are extracted using supervised and unsupervised methods, Conrad et al. [Conrad et al., 2009] concern themselves with the automatic generation of well-organised, fluent summaries of opinions about specified targets. The algorithm is provided with some opinion related questions about a specific target (see Table 2). The summarisation task is then the generation of summaries along the lines of the questionnaire. Identifying whether a sentence from a blog or review answers a certain target is done by modifying FastSum [Schilder and Kondadadi, 2008] (a summarisation method) slightly. It also uses a so-called cosine window function that assigns target score to words following the identified target. Even if the target is not explicitly present in some sentence, the sentence can still be considered.

**Table 3: Different types of relationships between aspects and opinions of a product from Somprasertsri and Lalitrojwong [Somprasertsri and Lalitrojwong, 2010]**

Relationship	Description
<i>child</i>	Aspect depends on the opinion. I <i>like</i> this <b>camera</b> .
<i>parent</i>	Opinion depends on the aspect. I have found that this camera take <i>incredible</i> <b>pictures</b> .
<i>sibling</i>	Both opinion and aspect depend on the same word. The <b>pictures</b> some time turn out <i>blurry</i> .
<i>grandchild</i>	Aspect depends on the word which depends on the opinion It's <i>great</i> having the <b>LCD display</b> .
<i>grandparent</i>	Opinion depends on the word which depends on the aspect. It has <b>movie mode</b> that works <i>good</i> for a digital camera.

More recently, the work of Somprasertsri and Lalitrojwong [Somprasertsri and Lalitrojwong, 2010] proposes an elaborate approach for extracting different aspects from an opinionated text fragments. Their work is most similar to that of Zhuang et al. [Zhuang et al., 2006] in the sense that it also tries to summarise opinions using opinion-aspect pairs and also make use of dependency grammar.

The try to model opinion-aspect pairs using different types of relationships (see Table 3). Their method first use different variations of noun phrases to extract possible candidates for aspects and then for each aspect candidate it finds the relevant opinion words. A probabilistic model is then used to predict the opinion-relevant product aspects. For alleviating the draw back of customers using different words to refer to the same aspect (e.g., memory card, compact flash, CF card and etc. for referring removable memory), they manually construct product ontologies through manufactures product description. Li et al. [Li et al., 2010] uses syntactic trees based on conditional random fields (CRF) to extract the dependencies between opinions and aspects and to pass the polarity information among different textual fragments through conjunctions.

## 7. Predicting Sentiment Orientation

For predicting the sentiment orientation of a sentence, Hu and Liu [2004] sum the individual orientation of each opinion word in a sentence. The sentence gets ranked according to the summed up orientation. If the orientation is neutral, they try to break the tie by adding the orientation of aspect words as well. In case of a tie once again, the sentence is assigned the orientation of the preceding sentence. If a negation word is close to some opinion word, the orientation is inverted.

In the method of Conrad et al. [Conrad et al., 2009] once the sentences that are relevant for the questionnaire have been identified, it uses a simple sentiment tagger, which is based on unigram lookup, for computing the sentiment orientation. Each token is queried to check if it is negative or positive and the polarity of a sentence is computed by subtracting negative tokens from the positive ones. If polarity is greater than 1 or less -1, the sentence is tagged positive or negative, respectively, otherwise it is tagged as neutral.

Blair-Goldensohn et al. [Blair-Goldensohn et al., 2008] additionally weigh each word with the likelihood of it being positive or negative. For each sentence  $x$  they calculate its  $RawScore(x) = \sum o_i$  by summing up the individual orientation  $o_i$  of each word  $w_i \in x$  and its  $Purity(x) = \frac{RawScore(x)}{\sum |o_i|}$ , where  $RawScore(x)$  is discounted by aggregated absolute orientation score of each word. These scores are used to rank the individual sentences under different settings. Top ranked sentences are then used in the summaries.

The methods of Zhuang et al. and Somprasertsri and Lalitrojwong [2006], Somprasertsri and Lalitrojwong [2010] extract aspect-opinion pairs. The polarity regarding each aspect is predicted using the opinion word it is paired. Both methods also take the negation word into consideration and change the polarity likewise. Li et al. [Li et al., 2010] extract the relevant opinions for each aspect from the opinion word that is nearest to the aspect. All these methods Zhuang et al. [2006], Somprasertsri and Lalitrojwong [2010], Li et al. [2010] generate summaries as a structured object aspect-opinion pairs.



---

## 8. Underlying Tools

---

There exist different kinds of text annotation tools for creating annotated corpora. CREAM or Magpie provide users with useful visual tools for manual annotation, web page navigation, reading semantic tags and browsing or provide infrastructure and protocols for manual stamping documents with semantic tags such as RDF annotation. Some of the semantic web based annotation tools include Protégé, OnTeA and GATE are introduced at the below gives more advantages.

### 8.1. Protégé

A number of semantic editors are available to work with OWL ontologies. Protégé is a popular ontology construction and annotation tool that is developed at Stanford University. Protégé supports the Web Ontology Language through an OWL Plugin, which allows a user to load an OWL ontology, annotate data and save annotation markup.

“Knowtator is a general-purpose text annotation tool that is integrated with the Protégé knowledge representation system. By modeling syntactic and/or semantic phenomena using Protégé frames, a wide variety of annotation schemas can be de-fined and used for annotating text. New annotation tasks can be created without writing new software or creating specialized configuration files. Know-tator also provides additional features that make it useful for real-world multi-person annotation tasks. Knowtator facilitates the manual creation of training and evaluation corpora for a variety of biomedical language processing tasks.”

### 8.2. OnTeA

Ontea is known as a Pattern based Semantic Annotation Platform. It works on text, in particular domain described by domain ontology and uses regular expression patterns for semi-automatic semantic annotation. Ontea tries to detect ontology elements within the existing application/domain ontology model. It means that by the Ontea annotation engine we can achieve the following objectives:

- Detecting Meta data from Text
- Preparing improved structured data for later computer processing
- Structured data are based on application ontology model

### 8.3. GATE

GATE is a framework for the development and deployment of language processing technology in large scale. GATE has several plugins useful for carrying out different types of information extraction tasks. It also has its own ontology API, which can be used together with these resources to take the maximum advantage of the combination. In traditional GATE is run over a corpus of texts to produce a set of annotated texts. In h-TechSight, the input to GATE takes the form of a set of URLs of target webpages, and an ontology of the domain. Its output comprises annotated instances of the concepts from the ontology. The ontology sets the domain structure and priorities with respect to relevant concepts with which the application is concerned. The GATE application consists of 5 basic stages:

1. web mining application to find relevant documents (or manual input of relevant documents)
2. selection of concepts in which the user is interested
3. information extraction
4. visual presentation of results (annotation of instances) and statistical analysis
5. ontology modification (an ontology editor is used to enrich the existing ontology from the results of the analysis)

## 9. Evaluation Metrics for Sentiment detection

The technical evaluation metrics includes factors related with the reliability (for example, the number of lost requests) and the usability (for example, average response time for each service request) on a mobile device. Behavioural metrics are related with the perceived efficiency in collecting citizen opinion, perceived transparency on the process making process and perceived improvement on citizen engagement and empowerment before and after the use of UbiPOL applications.

Table1: Metrics and Meanings

<b>Evaluication Metrics</b>	<b>Explanation</b>
The precision value	shows the proportion of correct opinions in all opinions
The recall value	gives the proportion of correct opinions in all correct opinions
F-measure	it is the harmonic mean of precision and recall:
Support value	represents the proportion of persons who give a positive score
Kappa score	represents topics and semantic category
Intensity value	gives range such as: weak-medium, strong
Polarity value	+, -, n
Conceptual similarity measure value	shows conceptual distance proposed by Wu et Palmer (Z.Wu and M.Palmer, 1994)
Characteristics of the reference corpus	Gives number of document, paragraph, sentence, word.

---

## 10. REFERENCES

---

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. ELRA. ISBN 29517408-6-7.

Albert Bifet and Frank Eibe. Sentiment knowledge discovery in Twitter streaming data. In *Proc 13th International Conference on Discovery Science*, Canberra, Australia, pages 1–15. Springer, 2010.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. Building a Sentiment Summarizer for Local Service Reviews. In *NLPIX*, 2008.

Stephan Bloehdorn, Philipp Cimiano, Andreas Hotho, Steffen Staab, An Ontology-based Framework for Text Mining, *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20, (1), Seiten 87-112, Mai, 2005.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

Jack G. Conrad, Jochen L. Leidner, Frank Schilder, and Ravi Kondadadi. Query-based opinion summarization for legal blog entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law, ICAIL '09*, pages 167–176, New York, NY, USA, 2009. ACM. ISBN 978-1-60558- 597-0.

Hang Cui, Vibhu Mittal, and Mayur Datar. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-2006)*, pages 1265–1270, 2006.

Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW '03*, pages 519–528, 2003.

Dmitry Davidov and Ari Rappoport. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44*, pages 297–304, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10*, pages 107–116, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-83-1.

Delmonte R., 2008. *Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution*, Nova Science Publishers, New York.

Domingue J., Dzbor M.: Magpie: supporting browsing and navigation on the semantic web. In IUI '04, pages 191-197, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-815-6.

Andrea Esuli and Fabrizio Sebastiani. 2006a. Determining term subjectivity and term orientation for opinion mining. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), pages 193–200, Trento, IT.

Andrea Esuli and Fabrizio Sebastiani. 2006b. SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pages 417–422, Genova, IT

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Pages 1–6, 2009.

Handschuh S., Staab S.: Authoring and annotation of web pages in cream. In WWW '02, pages 462-473, NY, USA, 2002. ACM Press. ISBN 1-58113-449-5. doi: <http://doi.acm.org/10.1145/511446.511506>

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining

Khelif, K., Dieng-Kuntz, R., Barbry, P.: An ontology-based approach to support text mining and information retrieval in the biological domain. Universal Computer Science 13(12), 1881–1907 (2007).

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: evaluating and learning user preferences. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, pages 514–522, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. Structure-aware review mining and summarization. In Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, pages 653–661, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

D. Maynard, A. Funk, and W. Peters. NLP-based support for ontology lifecycle development. In CK 2009 – ISWC Workshop on Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, Washington, USA, October 2009.

Raymond J. Mooney and R. Bunescu ,Mining Knowledge from Text Using Information Extraction SIGKDD Explorations (special issue on Text Mining and Natural Language Processing), 7(1):3-10, 2005.

Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM. ISBN 1-59593-046-9.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1.

Nitin Jindal and Bing Liu. 2006. Identifying Comparative Sentences in Text Documents. In Proc. of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval

Raymond J. Mooney, Razvan C. Bunescu: Mining knowledge from text using information extraction. SIGKDD Explorations 7(1): 3-10 (2005)

Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In K-CAP, pages 70–77, 2003.

Alexander Pak and Patrick Paroubek. Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, pages 436–439, Stroudsburg, PA, USA, 2010a. Association for Computational Linguistics.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2:1–135, January 2008. ISSN 1554-0669.

Pang, B., Lee, L., Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the ACL, pages 115–124, 2005

Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Turney, P.D., and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association, *ACM Transactions on Information Systems (TOIS)*, 21(4), 315-346.

Martin Potthast and Steffen Becker. Opinion summarization of web comments. In Advances in Information Retrieval, volume 5993 of Lecture Notes in Computer Science, pages 668–669. Springer Berlin / Heidelberg, 2010.

Horacio Saggion, Kalina Bontcheva, Adam Funk, and Diana Maynard, Ontology-based Information Extraction for Business Intelligence, Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference ISWC/ASWC2007, Busan, South Korea, 2007, Vol. 4825 Berlin, Heidelberg: Springer Verlag, November (2007), p. 837--850.

H. Saggion. Shaf: Semantic tagging and summarization techniques applied to crossdocument coreference. In Proceedings of SemEval 2007, Association for Computational Linguistics, pages 292, 295, 2007.

Frank Schilder and Ravikumar Kondadadi. Fastsum: fast and accurate query based multi-document summarization. In Proceedings of the 46th Annual Meeting of the Association for

Computational Linguistics on Human Language Technologies: Short Papers, HLT-Short '08, pages 205–208, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

G. Somprasertsri and P. Lalitrojwong. Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal Computer Science*, 16(6):938–955, 2010.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. Khoo. Sentiment classification of movie reviews using multiple perspectives. In *Proceedings of the 11<sup>th</sup> International Conference on Asian Digital Libraries: Universal and Ubiquitous Access to Information, ICADL 08*, pages 184–193, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-89532-9.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Demonstration Description in Conference on Empirical Methods in Natural Language Processing*, pages 34–35, 2005.

Turney, P.D., Coherent keyphrase extraction via Web mining, *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, 434–439.

P. Velardi, P. Fabriani, and M. Missikoff. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the ACM International Conference on Formal Ontology in Information Systems*, 2001.

Wilson, T., J. Wiebe, and R. Hwa, ‘Just How Mad Are You? Finding Strong and Weak Opinion Clauses’. In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004)*. San Jose, California, pp. 761–766.

René Witte and Christopher J.O. Baker. 2005. Combining Biological Databases and Text Mining to support New Bioinformatics Applications. In *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005)*, volume 3513 of LNCS, pages 310–321, Alicante, Spain, June 15–17.

C. Yang, K. H.-Y. Lin, and H.-H. Chen., Emotion Classification from Web Blog Corpora, *IEEE/WIC/ACM 2007*, 275–278.

Yunfang Wu and Peng Jin. Semeval-2010 task 18: Disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 81–85, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06*, pages 43–50, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2.