Automatic Sentiment Analysis in On-line Text

Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens

Katholieke Universiteit Leuven, Tiensestraat 41 B-3000 Leuven, Belgium
e-mail: erik.boiy@law.kuleuven.be; pieter.hens@econ.kuleuven.be
koen.deschacht@law.kuleuven.be; marie-france.moens@law.kuleuven.be

Abstract

The growing stream of content placed on the Web provides a huge collection of textual resources. People share their experiences on-line, ventilate their opinions (and frustrations), or simply talk just about anything. The large amount of available data creates opportunities for automatic mining and analysis. The information we are interested in this paper, is how people feel about certain topics. We consider it as a classification task: their feelings can be positive, negative or neutral. A sentiment isn't always stated in a clear way in the text; it is often represented in subtle, complex ways. Besides direct expression of the user's feelings towards a certain topic, he or she can use a diverse range of other techniques to express his or her emotions. On top of that, authors may mix objective and subjective information about a topic, or write down thoughts about other topics than the one we are investigating. Lastly, the data gathered from the World Wide Web often contains a lot of noise. All of this makes the task of automatic recognition of the sentiment in on-line text more difficult. We will give an overview of various techniques used to tackle the problems in the domain of sentiment analysis, and add some of our own results.

Keywords: sentiment analysis; document classification; artificial intelligence

1 Introduction

Automatic sentiment analysis is a topic within information extraction that only recently received interest from the academic community. In the previous decade, a handful of articles have been published on this subject. It's only in the last five years that we've seen a small explosion of publications. The idea of automatic sentiment analysis is important for marketing research, where companies wish to find out what the world thinks of their product; for monitoring newsgroups and forums, where fast and automatic detection of flaming is necessary; for analysis of customer feedback; or as informative augmentation for search engines.

The automatic analysis of sentiments on data found on the Web is useful for any company or institution caring about quality control. For the moment, getting user feedback means bothering him or her with surveys on every aspect the company is interested in. The problems with this approach are making a survey for each product or feature; the format, distribution and timing of the survey (asking to send a form right after purchase might not be very informative); and the reliance on the goodwill of people to take the survey. This method can be made obsolete by gathering such information automatically from the World Wide Web, where the large amount of available data creates the opportunity to do so. One of the sources are blogs (short for "web logs"), a medium through which the blog owner makes commentaries about a certain subject or talks about his or her personal experiences, inviting readers to provide their own comments. Another source are the electronic discussion boards, where people can discuss all kinds of topics, or ask for other people's opinions. We define a topic as the subject matter of a conversation or discussion, e.g. an event in the media or a new model of car, towards which the writer can express his or her views.

There are several additional advantages to this approach. First, the people who share their views usually have more pronounced opinions than average, which are additionally influencing others reading them, leading to so-called word-of-mouth marketing. Extracting these opinions is thus extra valuable. Second, opinions are extracted in real-time, allowing for quicker response times to market changes and for detailed time-based statistics that make it possible to plot trends over time.

This paper is organized as follows: In section 2 we will go over the concepts of emotions in written text. Section 3 gives an overview of various methods that can be used to analyse the sentiment of a text, making a distinction between symbolic techniques and machine learning approaches. In section 4 we describe some challenges in the field that need to be overcome. Section 5 provides a comparison of results from the literature using the aforementioned techniques, to which we add some of our own results1. In section 6 we shortly discuss those results, before coming to conclusions in section 7.

2 Concepts of Emotions in Written Text

2.1 Concept of Emotions

Before attempting to classify sentiments, we must ask the question what sentiments are. In general we can state that sentiments are either emotions, or they are judgements or ideas prompted or coloured by emotions[2]. An emotion consists of a set of stages, namely: appraisal, neural and chemical changes and action readiness. We will give a quick overview of each of these states.

An emotion is usually caused by a person consciously or unconsciously evaluating an event, which is denoted *appraisal* in psychology. Appraisal does not only denote the evaluation whether something is positive or negative, but it also denotes other measurements such as the significance of an event, the personal control or the involvement of the own ego. In general, the same appraisal gives rise to the same emotion. Appraisal causes *mental and bodily changes*, that make up the actual experience of an emotion. Emotions urge for actions and prompt for plans: an emotion gives priority for one or a few kind of *actions* to which it gives a sense of urgency. We use the term "action" to denote all mental or physical actions (that are the result of an emotion). This includes actions such as moving away from a negative event, mental processes, such as worrying about the event, and other effects that are direct result of the emotion, such as crying or going pale.

2.2 Emotions in Written Text

The study of emotions in text can be conducted from two points of view. Firstly, one can investigate how emotions influence a writer of a text in choosing certain words and/or other linguistic elements. Secondly, one can investigate how a reader interprets the emotion in a text, and what linguistic clues are used to infer the emotion of the writer. In this text, we'll take the second point of view. We are interested in the way people infer emotions, so we can mimic this process in a computer program. In the remainder of this section we will investigate how linguistic elements describing appraisal and action-readiness are used in texts to convey the emotion of the author, as they comprise the majority of clues to infer emotion from text.

Appraisal

A lot of linguistic scholars agree on the three dimensions of Osgood and al. [1], who investigated how the meaning of words can be mapped in a semantic space. Factor analysis extracted 3 major dimensions: (1) positive or negative evaluation (2) a power, control or potency dimension and (3) an activity, arousal or intensity dimension. Although these dimensions are originally proposed as the dimensions of a semantic space, they can also be used to organize linguistic categories of emotion or for the automatic detection of emotions. Most research is devoted towards the appraisal component of emotions, and we will look into it a bit deeper by briefly going over Osgood's dimensions, giving some examples along the way.

(1) Evaluation (positive/negative)
The evaluation dimension is fairly straightforward; it contains all choices of words, parts of speech, word organization patterns, conversational techniques, and discourse strategies that express the orientation of the writer to the current topic. Evaluation is often expressed by using adjectives.
e.g. “It was an *amazing* show.”

(2) Potency (powerful/unpowerful)
This dimension contains all elements that general express whether the writer identifies and commits himself towards the meaning of the sentence or whether he dissociates himself. From a psychological standpoint these phenomena are related to approach and avoidance behaviour. This dimension consists of 3 sub-dimensions: proximity, specificity and certainty.

(2.1) Proximity (near/far)

[1] This research is sponsored by the company Attentio (Brussels, Belgium): attentiocontact@attentio.com. We are very grateful to Per Siljubergsåsen and Simon Mc Dermott for the funding and the relevant remarks. We thank Kalina Lipinska for the annotation of one of the corpora we use in our experiments. We thank Roxana Angheluta for her valuable help when integrating our work in Attentio’s system architecture.

[2] Adapted from the Merriam-Webster On-line Search dictionary.

This category contains all linguistic elements that indicate the 'distance' between the writer and the topic. The proximity from the writer to the current topic expresses whether the writer identifies himself with the topic or distance himself from it.
e.g. “I'd like you to meet John.” versus “I'd like you to meet Mr. Adams.” (social proximity)

(2.2) Specificity (clear/vague)
Specificity is the extent to which a conceptualized object is referred to by name in a direct, clear way; or is only implied, suggested, alluded to, generalized, or otherwise hinted at.
e.g. “I left *my* / *a* book in your office.” (particular vs general reference)

(2.3) Certainty (confident/doubtful)
This dimension expresses the certainty of the writer towards the expressed content. A stronger certainty indicates that the writer is entirely convinced about the truth of his writings and possibly indicates a stronger emotion.
e.g. “It *supposedly* is a great movie.” versus “It *definitely* is a great movie.”

(3) Intensifiers (more/less)
When expressing emotions, a lot of the emotional words used do not express an emotion, but modify the strength of the expressed emotion. These words, the intensifiers, can be used to strengthen or weaken both positive and negative emotions.
e.g. “This is *simply* the best movie.” (adverb)
“He had cuts *all* over.” (quantifier)
“Where *the hell* have you been?” (swearing)

Direct Expressions

The most direct way to express an emotion is of course to express it directly, without making a detour by using appraisal or action readiness. This can be done among others by using verbs and adjectives [2, 3]. A typical way to express an emotion directly seems to be a pattern similar to "I am/feel/seem [adjective describing emotion]"
e.g. I *ache for* a cigarette.
I *am delighted* of the final results.

Elements of Action

Excellent examples of actions indicating emotions are of course crying and laughing, but more subtle signs that denote emotion in certain circumstances can be considered as well. An example is looking at your watch when watching a movie, which is most probably a result of boredom and a lack of interest.
e.g. I was *grinning* the whole way through it and *laughing out loud* more than once.

Remarks

There are additional ways of expressing emotions that don't strictly fall into above categories, like the use of figurative language and irony. It must also be noted that most techniques in sentiment classification focus on terms that do actually not really denote emotions, but denote evaluation, appreciation or judgement. Of course this is not surprising, because most techniques focus on reviews of movies, products, cars, etc., and basically in a review the reviewer evaluates the object under discussion. The sentiment of the reviewer is often not discussed, although of course, it is often easy to infer his emotions. Recognizing the fact that classifying a review is in essence classifying it according to appraisal, doesn't only improve understanding but can also lead to the discovery of new techniques.

3 Methodology

In the previous section we discussed the indicators of sentiment in text. In this section we will see methods of identifying this information in a written text. There are two main techniques for sentiment classification: symbolic techniques and machine learning techniques. The symbolic approach uses manually crafted rules and lexicons, where the machine learning approach uses unsupervised, weakly supervised or fully supervised learning to construct a model from a large training corpus.

3.1 Symbolic Techniques

3.1.1 Lexicon Based Techniques

The simplest representation of a text is the bag-of-words approach. Hereby, we simply consider the document as a collection of words without considering any of the relations between the individual words. Next, we determine the sentiment of every word and combine these values with some aggregation function (such as average or sum). We will discuss a selection of methods to determine the sentiment of a single word.

3.1.1.1 Using a Web Search

It was already indicated by Hatzivassiloglou and Wiebe [4] that adjectives are good indicators of subjective, evaluative sentences. Turney [5] recognizes that, although an isolated adjective may indicate subjectivity, there may be insufficient context to determine semantic orientation. For example, the adjective "unpredictable" may have a negative orientation in an automotive review, in a phrase such as "unpredictable steering", but it could have a positive orientation in a movie review, in a phrase such as “unpredictable plot". Therefore he used tuples consisting of adjectives combined with nouns and of adverbs combined with verbs.

The tuples are extracted from the reviews and the semantic orientation of a review is calculated as the average semantic orientation of the tuples taken from that review. To calculate the semantic orientation for a tuple (such as "unpredictable steering"), Turney uses the search engine Altavista. For every combination, he issues two queries: one query that returns the number of documents that contain the tuple close (defined as “within 10 words distance”) to the word "excellent" and one query that returns the number of documents that contain the tuple close to the word "poor". If the combination is found more often in the same context as "excellent" than in the same context as "poor", the combination is considered to indicate a positive orientation, and otherwise to indicate a negative orientation.

3.1.1.2 Using WordNet

Kamps and Marx use WordNet [6] to determine the orientation of a word. In fact, they go beyond the simple positive/negative orientation, and use the dimension of appraisal that gives a more fine-grained description of the emotional content of a word. Kamps and Marx developed an automatic method [7] using the lexical database WordNet to determine the emotional content of a word along Osgood et al.'s dimensions. In essence, the WordNet database consists of nodes (the words) connected by edges (synonym relations). Kamps and Marx define a distance metric between the words in WordNet, called minimum path-length (MPL). This distance metric counts the number of edges of the shortest path between the two nodes that represent the words. For example, the words "good" and "big" have a MPL of 3. The shortest path from the word "good" to the word "big" is the sequence <good, sound, heavy, big>.

To estimate the magnitude of a dimension of appraisal for a particular word, they compare the MPL of that word towards the positive and towards the negative end of that dimension. Both ends of a dimension are represented by prototype-words. The positive end of the evaluative dimension is represented by the word "good" and the negative end is represented by the word "bad". The prototypes for the potency dimension are respectively "strong" and "weak" and for the activity dimension "active" and "passive".

Only a subset of the words in WordNet can be evaluated using this techniques, because not all words are connected to one of the prototype words. After examination, it showed that the subset of words connected to either "good" or "bad" was composed of 5410 words. Interestingly, the subset of words connected to either "strong" or "weak" consisted of exactly the same 5410 words, and so did the subset connected to "active" or "passive". It seems that all important words expressing emotive or affective meaning are included in this one set.

3.1.2 Sentiment of Sentences

So far, we've seen different methods that determine the sentiment of a single word and assumed a simple approach to combine the sentiments of words within a single sentence. The bag-of-words approach has some important drawbacks. As already briefly indicated in section 3.1.1.1, it can often be advantageous to consider some relations between the words in a sentence. There are several approaches in this field; we mention here briefly Mulder and al.'s article [8], which discusses the successful use of an affective grammar. They note that simply detecting emotion words can tell whether a sentence is positive or negative oriented, but does not explain towards what topic this sentiment is directed. In other words, what is lacking in the research towards affect is the relation between attitude and object. Mulder and al. have studied how this relation between attitude and object can be formalized. They combined a lexical and grammatical approach: (1) lexical, because they believe that affect is primarily expressed through affect words, and (2) grammatical, because affective meaning is intensified and propagated towards a target through function-words and grammatical constructs.

3.2 Machine Learning Techniques

In this section a description and comparison of state-of-the-art machine learning techniques used for sentiment classification are discussed. First a description is given of a selection of different features that are commonly used to represent a document for the classification task, followed by an overview of machine learning algorithms.

3.2.1 Feature Selection

The most important decision to make when classifying documents, is the choice of the feature set. Several features are commonly used, like unigrams or part-of-speech (the linguistic category of a word, further shortened to "POS") data. Features and their values are commonly stored in a feature vector.

Unigrams
This is the classic approach to feature selection, in which each document is represented as a feature vector, where the elements indicate the presence (or frequency) of a word in the document. In other words, the document is represented by its keywords.

N-grams
A word N-gram is a subsequence of N words from a given sequence (e.g. a sentence). This means that the features in the document representation are not single words, but pairs (bigrams), triples (trigrams) or even bigger tuples of words. For example, "easy" followed by "to" becomes "easy to" in a bigram. Other examples of positive oriented bigrams are: "the best", "I love", "the great", ... and negative oriented: "not worth", "back to", "returned it", ... [9]. With the use of N-grams it is possible to capture more context. N-grams are for example effective features for word sense disambiguation [10]. When using N-grams, the feature vector could take on enormous proportions (in turn increasing sparsity the of the feature vectors). Limiting the feature vector size can be done by setting a threshold for the frequency of the N-grams, or by defining rule sets (e.g. only incorporate N-grams that satisfy a certain pattern like *Adjective Noun* or *Adverb Verb*).

Lemmas
Instead of using the words as they literally occur in the text, the lemmas of these words can be used as features for the document. This means that for each word its lemma, being its basic dictionary form, is identified. Examples are:

writes -> write *was -> be* *better -> good*
written -> write *cars -> car* *best -> good*

The advantage with lemmatisation is that the features are generalized and it will be easier to classify new documents, but this is not always true: you still have to look out for overgeneralization. Dave et al. [9] report a decrease in accuracy of sentiment classification when the words in the documents are conflated to their dictionary form. Lemmatisation comes with loss of detail in the language. For example, Dave notes that negative reviews tend to occur more in the past tense, which cannot be detected after lemmatisation.

Negation
Another extension of the unigram approach is the use of negation. When you only consider the words in a sentence and someone writes "*I don't like this movie",* a program can think that this person loved the movie, when it looks at the word "like". A solution for this is to tag each word after the negation until the first punctuation (with for example NOT_). The previous sentence will then become: "*I don't NOT_like NOT_this NOT_movie"*. This was done by [11]. In this experiment, the negation tagging gives a slight decrease in performance. Dave et al. [9] note that simple substrings (N-grams) work better at capturing negation phrases.

Opinion Words
Opinion words are words that people use to express a positive or negative opinion [12]. Opinion words are obtained from several POS classes: adjectives, adverbs, verbs and nouns [13, 14]. These opinion words can be

incorporated into the feature vector, where they represent the presence or absence of such a word. Two techniques can be used to define opinion words:

- Use a predefined lexicon; Wiebe and Riloff [14] constructed such an opinion word-list. This approach combines the lexicon based method described above with the machine learning methods.
- Identify the words (mostly adjectives; see below) that describe a certain feature of a product in a text [12]. e.g. After nearly 800 pictures I have found that this camera takes *incredible* pictures.

Adjectives

Wiebe noted in [15] that adjectives are good indicators for subjectivity in a document. According to these findings you can assume that documents only represented by their adjectives should do well in sentiment classification. Experiments where only adjective features are used, were done in [11, 16]. The results showed that you get better results when using all POS data. This doesn't mean that adjectives are bad sentiment classifiers, as adjectives only represent on average 7.5% of the text in a document.

Salvetti used WordNet to enrich the only-adjective feature vectors. He translated the adjectives into synsets of adjectives and used hypernym generalization on them (both synsets and hypernyms can be found using WordNet). Using this procedure he found a decrease in the accuracy of the sentiment classification, which was due to the loss of information produced by the generalization.

3.2.2 Machine Learning Techniques

Supervised Methods

In order to train a classifier for sentiment recognition in text, classic supervised learning techniques (e.g. Support Vector Machines, naive Bayes Multinomial, Maximum Entropy) can be used. A supervised approach entails the use of a labelled training corpus to learn a certain classification function. The method that in the literature often yields the highest accuracy regards a Support Vector Machine classifier [11]. In the following section we discuss a selection of classification algorithms. They are the ones we used in our experiments described below.

(1) Support Vector Machines (SVM)

Support Vector Machines operate by constructing a hyperplane with maximal Euclidean distance to the closest training examples. This can be seen as the distance between the separating hyperplane and two parallel hyperplanes at each side, representing the boundary of the examples of one class in the feature space. It is assumed that the best generalization of the classifier is obtained when this distance is maximal. If the data is not separable, a hyperplane will be chosen that splits the data with the least error possible.

(2) Naive Bayes Multinomial (NBM)

A naive Bayes classifier uses Bayes rule (which states how to update or revise believes in the light of new evidence) as its main equation, under the naive assumption of conditional independence: each individual feature is assumed to be an indication of the assigned class, independent of each other. A multinomial naive Bayes classifier constructs a model by fitting a distribution of the number of occurrences of each feature for all the documents.

(3) Maximum Entropy (Maxent)

The approach tries to preserve as much uncertainty as possible. A number of models are computed, where each feature corresponds to a constraint on the model. The model with the maximum entropy over all models that satisfy these constraints is selected for classification. This way no assumptions are made that are not justified by the empirical evidence available.

Unsupervised and Weakly-supervised Methods

The above techniques all require a labelled corpus to learn the classifiers. This is not always available, and it takes time to label a corpus of significant size. Unsupervised methods can label a corpus, that is later used for supervised learning (especially semantic orientation is helpful for this [17]). Turney's technique using AltaVista (see section 3.1.1.1) can be viewed as a form of weakly supervised learning, where a set of seed terms is expanded to a collection of words. We mention two more methods for determining the sentiment of single words based on weakly-supervised methods. Hatzivassiloglou and McKeown[18] presented a method for determining the sentiment of adjectives by clustering documents into same-oriented parts, and manually label the clusters positive or negative. OPINE [19] is a system that uses term clustering for determining the semantic orientation of an opinion word in combination with other words in a sentence. The idea behind this approach comes from the fact that the orientation of a word can change with respect to the feature or sentence the word is associated (e.g. The word *hot* in the pair: *hot water* has a positive sentiment, but in the pair *hot room* it has a negative sentiment).

4 Challenges

With the techniques described above, pretty good results can be obtained already (see section 5), but nevertheless, there are some challenges that need to be overcome.

4.1 Topic-Sentiment Relation

Our goal is to determine sentiments towards a certain topic. It often happens that a person expresses his opinion towards several topics within the same text or sentence. For example, in a movie review he may state he dislikes the special effects and some of the acting, but likes the movie nonetheless. His opinion about these topics is in contradiction with his thoughts about the movie in general. When a sentence contains a lot of negative subjectivity, but all expressed toward a different topic than the one we are investigating, the sentence is still classified as negative. Therefore, it is useful to investigate the relation of the sentiment to the topic. One way of doing this is by looking into the sentence parse tree (i.e. a syntactic analysis of the sentence according to the language's grammar) to derive better features.

Related to this problem is the classification of whole texts. Until now we have only looked at the classification of sentences, in which topic and terms indicative for the sentiment are assumed to appear together. This is however not a realistic assumption. For the detection of the topic-sentiment relation in texts, coreference resolution needs to be applied across sentences. Even when there is only one topic in the text, it is also advantageous to use a more advanced metric to combine the predictions for the sentences than a simple sum of the sentiments found in the individual sentences. Taboada and Grieve [20] state that opinions expressed in a text tend to be found in the middle and the end of that text. Therefore, they weigh the semantical orientation of a sentence based on its position in the text, giving improved results.

4.2 Neutral Text

A first question is what to do with neutral text, as not all text is either positively or negatively oriented. It is often useful to determine whether a piece of text expresses subjective or objective content. Subjective sentences are used to communicate the speaker's evaluations, opinions, emotions and speculations, while objective sentences are used to convey objective, factual information [21]. Both kinds often appear in the same text, for instance in movie reviews, where the writer can express his attitude toward the movie (which is the semantic orientation of the document), but can also describe, within the same review, objective statements about the movie itself (e.g. a summary of the plot). Most subjectivity classifiers use machine learning techniques (see [22]) and classify between subjective and objective sentences or between positive, negative and objective sentences. To our best knowledge, there has been only one attempt to use a symbolic technique that classifies subjective sentences, done by Wiebe [23, 24].

4.3 Cross-domain Classification

Other research in the sentiment classification field regards cross-domain classification. How can we learn classifiers on one domain and use them on another domain (e.g. *books* and *movies*)? A reason why cross-domain classification might be necessary is because there is not always enough training data available to train a classifier for a specific domain. The classifier should then be trained with data from another domain. Tests are done by Aue et al. [25] and by Finn et al. [26]. Overall, they show that sentiment analysis is a very domain-specific problem, and it is hard to create a domain independent classifier. One possible approach is to train the classifier on a domain-mixed set of data instead of training it on one specific domain.

4.4 Text Quality

A last important issue is the quality of the text to be evaluated. When text is automatically gathered from the World Wide Web, one can expect a fair amount of junk to be returned (e.g. adds, web site menus, links, ...). This junk may be mixed with other information we are interested in, making it more difficult to filter it out. Also the language used by the writers may be of poor quality, containing lots of Internet slang and misspellings. Both issues have a negative influence on the classification for both types of methods discussed. Especially the junk may confuse a machine learner by providing it with a lot of irrelevant features. This also means extensive manual filtering of the text in order to acquire a good training corpus, and makes it harder to perform deeper

NLP techniques like parsing. An example of dirty input text (the topic is the movie “A Good Year”) is the following:

Nothing but a French kiss-off Search Recent Archives Web for (rm) else &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; &#8226; ONLINE EXTRAS SITE SERVICES Movie Listings Friday Nov 10 2006 Posted on Fri Nov. 10 2006 MOVIE REVIEW A Good Year a flat bouquet Nothing but a French kiss-off Gladiator collaborators seem defeated by light-weight love story.By ROBERT W.

Needless to say that using a sentence parser (that detects the syntactic structure of the sentence) on this example will have little success.

5. Results

5.1 Evaluation Measures

As a first evaluation measure we simply take the classification accuracy, meaning the percentage of examples classified correctly. This measure is not sufficient when we classify individual sentences and include the neutral class as a third option next to the positive and negative ones. Neutral examples are a majority in texts and their correct detection largely influences simple accuracy results. With this in mind, it makes more sense for us to use precision for positive and negative examples as the evaluation measure. When generating sentiment statistics a high recall is just as desirable as a high precision. Other evaluation metrics that influence the performance are also considered: the speed of the classification method, the feature vector size and the available resources.

5.2 Symbolic Techniques

In the previous section we have seen a selection of two symbolic techniques, for which we give concrete results. Turney reports accuracies ranging between 65.83% on a collection of movie reviews, to 84.0% on a collection of automobile reviews when applying his method using a web search engine. Kamps and Marx achieved an accuracy of 68.19% on the manually constructed list of the General Inquirer for classification along Osgood's evaluative dimension, when applying their approach using WordNet. The accuracy rose to 76.72% when increasing the interval for which words are considered neutral.

An interesting experiment done by Pang et al. [11] shows the difficulty to construct a lexicon (or another knowledge-resource) that has a (close to) complete coverage of the target domain. A lot of information is often not captured in the hand-built model and lost. In the experiment they compared the ability of humans in selecting appropriate words for an emotion lexicon, with automatic methods. Although the lists of words created by humans seemed intuitively valid, they resulted in poorer performance: the best human created list resulted in 64% accuracy (with 39% ties), while a simple automatic method (a count of the frequencies of words in positive and negative reviews) resulted in a list with 69% accuracy and only 16% ties. Interestingly, some words that have no significant emotional orientation were quite good indexes. For example, the word “still", was found to be a good indicator of positive orientation, because it appeared in sentences such as “Still, though, it was worth seeing".

Given the above results, we did not perform any experiments with symbolic techniques, instead we focused on machine learning techniques of which the results are given below.

5.3 Machine Learning Techniques

5.3.1 Corpora

A corpus is a large, electronically stored set of texts. Corpora are used by machine learning approaches both for training and testing (and just for testing in the case of symbolic approaches). Evaluation will often be performed by using cross-validation. This means that over several iterations, in each iteration part of the corpus will be used for training, and the other part for testing. After all iterations, each example from the corpus will have been used for testing once, resulting in a full evaluation of the corpus. In order to compare results of different approaches, they need to be compared on the same corpus, as some corpora can be considerably easier to work with than others. We performed tests on two corpora to obtain the results presented in this paper:

- Pang and Lee's[3] movie review corpus, consisting of 1000 positive and 1000 negative reviews, is often used to evaluate sentiment analysis approaches in the literature. These movie reviews seem hard to classify. A possible explanation of this phenomenon is the mix of words that describe the storyline and words that describe the evaluation of that particular movie.
- A corpus gathered from blogs, discussion boards and other websites containing 759 positive, 205 negative, 1965 neutral and 1562 junk examples, annotated with a sentiment towards the topic under evaluation. The latter two categories were considered as one for our test purposes. The topics include various movie titles and car brands. The examples are of poor quality, displaying the problems described in section 4.4 (the example given there was taken from this corpus). As the number of examples in each category is very unbalanced, corrective measures were taken by adding additional examples from the Customer Review Datasets corpus by Hu and Liu[4]. In total, 550 negative sentences from the customer reviews were added to the corpus, and 222 extra positive sentences were used for training only.

5.3.2 Our Experiments

In Table 1 we show some of our results on the movie review corpus, indicating the features that perform well in the literature (discussed above), optional processing and the machine learning methods used. For both the support vector machine (SVM) and naive Bayes multinomial (NBM) methods the Weka[5] implementation was used, the Maxent[6] package from OpenNLP was used as implementation of the maximum entropy classifier. For our tests using SVM's, an error tolerance of 0.05 was set for training, the other parameters (e.g. linear kernel) were kept default for all methods. We used QTAG as POS tagger for obtaining the adjectives. It achieves a rather low accuracy[7], but it is fast and easy to incorporate into software. “Subjectivity analysis” stands for a simple subjectivity analysis using a NBM classifier, trained on the subjectivity dataset introduced in Pang and Lee [22], which removes all objective sentences from the examples. A cut-off of four was used for the bigram feature, meaning that only bigrams occurring at least four times were included in the feature vector. Frequencies of the features were used in the feature vector for SVM and NBM, while binary feature presence was used for Maxent.

Features	***SVM***	***NBM***	***Maxent***
Unigrams	85.45%	81.45%	84.80%
Unigrams & subjectivity analysis	86.35%	83.95%	87.40%
Bigrams	85.35%	83.15%	85.40%
Adjectives	75.85%	82.00%	80.30%

Table 1: Results in terms of accuracy on the movie review corpus for different machine learning methods using a selection of features (and processing)

Table 2 shows our results on the second corpus that realistically represents blogs found on the World Wide Web. The corpus was extended with 550 negative review sentences, which are included in the results. In the first column are the baseline results on the corpus. The baseline uses the approach that gives the best results for the movie corpus (see Table 1), i.e., an approach comparable to the literature and with a low novelty factor. In the second column are our latest results. A total of 84 examples were beyond the reach of our current methods and are excluded from the results. In order to include those examples, we could consider them as neutral; resulting in a slight decrease in the total accuracy and in the recall for positive and negative, compared to the results shown in the second column, while still being much better than the baseline results. The methods, features and processing used to arrive to these results may not be disclosed by us. For more information on the methods used, the reader may contact Attentio, the company that sponsors our research.

[3] Available at http://www.cs.cornell.edu/people/pabo/movie-review-data.

[4] Available at http://www.cs.uic.edu/~liub/FBS/FBS.html.

[5] See http://www.cs.waikato.ac.nz/~ml/weka/.

[6] See http://maxent.sourceforge.net/.

[7] Our own experiments indicate an accuracy of about 86%, while current state of the art POS tagging achieve ca. 96% accuracy.

	Baseline NBM	***Our latest approach***
accuracy %	84.25	90.25
precision/recall % for positive	64.52/49.93	74.39/75.62
precision/recall % for negative	88.48/72.96	87.43/82.70

Table 2: Results on the blog corpus, comparing the results of the baseline approach (cf. Table 1) and those of our latest methods

6 Discussion

Although we have not done any experiments using symbolic techniques ourselves, we deemed machine learning approaches more promising after reviewing methods from both categories, and conducted our research in that direction. Judging from the good results we have achieved, this seems like it has been the right choice.

The results in Table 1 show that there is rather little difference in accuracy between the experiments using different features (except for the adjectives). With this in mind, it becomes interesting to look at other factors influencing the choice of which features and processing to use. The advantages of unigrams and bigrams over the other features are that they are faster to extract, and require no extra resources to use, while e.g. adjectives require a POS tagger to be run on the data first, and subjectivity analysis requires an additional classifier to be used. A downside is the feature vector size, which is substantially (over 5 times for unigrams) larger e.g. than when only adjectives are included. For the machine learning method we see a more substantial difference between NBM and both SVM and Maxent. It might however still be advantageous to use NBM, as it is considerably faster. The results of the state of the art techniques for sentiment classification on the movie review corpus shown in Table 1 are comparable with the ones found in the literature that use this corpus.

The results from Table 2 need some more explanation. The blog corpus used in the experiments of Table 2 is considerably more difficult to work with, and is annotated in three classes (including neutral), where the movie review corpus (results in Table 1) only had two. However, compared to the baseline (current state-of-the-art) approach, our latest method performs significantly better. The lower precision and recall for the positive class compared to the negative one, are due to the added negative examples from the easier Customer Review Datasets corpus, and due to the higher correlation of positive examples with neutral ones, making misclassifications between those classes more common. The results we obtained are encouraging, and show that it is possible to overcome the difficulties explained in section 4.

7 Conclusion

In this paper we have indicated the usefulness of sentiment classification, and have given an overview of the various methods used for this task. While many of the methods show encouraging results, there are still challenges to be overcome when applying them to data gathered from the World Wide Web, especially from blogs. We have demonstrated that in these circumstances improvements over state of art methods for sentiment recognition in texts are possible.

References

[1] OSGOOD, C. E.; SUCI, G. J; TANNENBAUM, P. H. *The Measurement of Meaning*. University of Illinois Press, 1971 [1957].

[2] BIBER, D; FINEGAN, E. Styles of stance in english: *Lexical and grammatical marketing of evidentiality and affect*. Text 9, 1989, pp. 93-124.

[3] WALLACE, A. F. C.; CARSON, M. T., *Sharing and diversity in emotion terminology*. Ethos 1 (1), 1973, pp. 1-29.

[4] HATZIVASSILOGLOU, V.; WIEBE, J., *Effects of adjective orientation and gradability on sentence subjectivity*, Proceedings of the 18th International Conference on Computational Linguistics, ACL, New Brunswick, NJ, 2000.

[5] TURNEY, P., *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*, Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[6] FELLBAUM, C. (ed.), *Wordnet: An electronic lexical database*, Language, Speech, and Communication Series, MIT Press, Cambridge, 1998.

[7] KAMPS, J.; MARX, M.; MOKKEN, R. J.; DE RIJKE, M., *Using WordNet to measure semantic orientation of adjectives.* LREC 2004, volume IV, pp. 1115—1118.

[8] MULDER, M.; NIJHOLT, A.; DEN UYL, M.; TERPSTRA, P., *A lexical grammaticaimplementation of affect*, Proceedings of TSD-04, the 7th International Conference Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 3206, Springer-Verlag, Brno, CZ, 2004, pp. 171–178.

[9] DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.* In Proceedings of WWW-03, 12th International Conference on the World Wide Web, ACM Press, Budapest, HU, 2003, pp. 519–528.

[10] PEDERSEN, T. *A decision tree of bigrams is an accurate predictor of word sense.* In Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 2001, pp. 79–86.

[11] PANG, B.; LEE, L.; VAITHYANATHAN, S. *Thumbs up? Sentiment classification using machine learning techniques.* In Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Philadelphia, US, 2002, pp. 79–86.

[12] HU, M.; LIU, B. *Mining opinion features in customer reviews.* In Proceedings of AAAI-04, the 19th National Conference on Artificial Intellgience, San Jose, US, 2004.

[13] BETHARD, S.; YU, H.; THORNTON, A.; HATZIVASSILOGLOU, V.; JURAFSKY, D. *Automatic extraction of opinion propositions and their holders.* In James G. Shanahan, Janyce Wiebe, and Yan Qu, editors, Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004.

[14] RILOFF, E.; WIEBE, J.; WILSON, T. *Learning subjective nouns using extraction pattern bootstrapping.* In Walter Daelemans and Miles Osborne, editors, Proceedings of CONLL-03, 7th Conference on Natural Language Learning, Edmonton, CA, 2003, pp. 25–32.

[15] WIEBE, J. *Learning subjective adjectives from corpora.* In Proceedings of AAAI-00, 17th Conference of the American Association for Artificial Intelligence, AAAI Press / The MIT Press, Austin, US, 2000, pp. 735–740.

[16] SALVETTI, F.; LEWIS, S.; REICHENBACH, C. *Impact of lexical filtering on overall opinion polarity identification.* In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004.

[17] BEINEKE, P.; HASTIE, T.; VAITHYANATHAN, S. *The sentimental factor: Improving review classification via human-provided information.* In Proceedings of ACL-04, the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, ES, 2004, pp. 263–270.

[18] HATZIVASSILOGLOU, V.; MCKEOWN, K. R. *Predicting the semantic orientation of adjectives.* In Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Madrid, ES, 1997, pp. 174–181.

[19] POPESCU, A.; ETZIONI, O. *Extracting product features and opinions from reviews.* In Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing, Vancouver, CA, 2005.

[20] TABOADA, M.; GRIEVE, J. *Analyzing appraisal automatically.* In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, 2004, pp. 158–161.

[21] WIEBE, J.; BRUCE, R. F.; O'HARA, T. P. *Development and use of a gold-standard data set for subjectivity classifications.* In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, College Park,US, 1999, pp. 246–253.

[22] PANG, B.; LEE, L. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* In Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Barcelona, ES, 2004, pp. 271–278.

[23] WIEBE, J. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text.* Technical report, SUNY Buffalo Dept. Of Computer Science, Buffalo, NY, 1990.

[24] WIEBE, J. *Tracking point of view in narrative.* Computational Linguistics, 20 (2), 1994, pp. 233–287.

[25] AUE, A.; GAMON, M. *Customizing sentiment classifiers to new domains: a case study.* In Submitted to RANLP-05, the International Conference on Recent Advances in Natural Language Processing, Borovets, BG, 2005.

[26] FINN, A.; KUSHMERICK, N. *Learning to classify documents according to genre.* J. American Society for Information Science and Technology, Special issue on Computational Analysis of Style, 57(9), 2006.