

# Sentiment Analysis in Parliamentary Proceedings

## Steven Grijzenhout, Maarten Marx, Valentin Jijkoun

### Abstract

The question is addressed whether opinion mining techniques can successfully be used to automatically retrieve political viewpoints from parliamentary proceedings. Two specific preprocessing tasks are identified and systematically evaluated: automatically determining subjectivity in the publications and automatically determining the semantic orientation of the subjective parts. A corpus of recent parliamentary proceedings has been collected and a gold standard annotation is created on both subjectivity and orientation. Following this, a number of models based on subjectivity lexicons and machine-learning algorithms are evaluated. Machine-learning algorithms perform best, but methods based on subjectivity lexicons also provide promising results. Based on these results we can conclude that opinion mining techniques applied to political data score just as well as the state of the art in other more traditional domains of opinion mining like product reviews and blogs.

**Keywords:** opinion mining, subjectivity, semantic orientation, parliamentary proceedings

## 1 Introduction

Opinion mining is a recent discipline concerned with automatically determining the opinion a text expresses (Pang & Lee, *Opinion Mining and Sentiment Analysis*, 2008). Opinion mining and sentiment analysis are terms that are used more or less synonymously in the literature, and we will also use them interchangeably. In this paper, we evaluate whether opinion mining techniques can successfully be applied in political text analysis. As data we use the verbatim transcriptions of the plenary meetings in the Dutch House of Representatives. These documents are an important source of information on the position of political parties and individuals in the political arena. Our research concerns the following central research question: Can opinion-mining techniques be used to automatically retrieve political viewpoints from parliamentary proceedings?

To answer this question (albeit indirectly), we evaluate whether opinion-mining techniques are appropriate methods to analyse political data. This approach is warranted as opinion-mining techniques have been tested elaborately on product reviews and blogs, but not yet on political data. We will show that the data format of parliamentary proceedings is well suited to do sentiment analysis. We then proceed to evaluate the two key steps in sentiment analysis: determining subjective passages, and in these passages, determine the semantic orientation, that is: do they express positive or negative attitude?

The paper is organized as follows: Section 2 provides background information on the various approaches to sentiment analysis; Section 3 describes the data; Section 4 describes the task and the experiments we have done, and presents the results; Section 5 reviews methods and results and compares them with other research in this field.

## 2 Background

This research is part of a set of four research areas: opinion search, opinion mining, topic mining and recent research regarding Dutch parliamentary proceedings. In this section these research areas are discussed briefly.

*1. Opinion search.* Opinion search is a relatively new branch of research. That aims to enable users to search for opinions on any object (Liu, 2007). Here, the entity “object” is used to point to different concepts including products, persons, happenings or topics. Therefore opinion search can be helpful for a broad range of applications, including review-related websites, blogs, business intelligence, government intelligence and politics. Most

research covers opinion search applications in the context of blogs (web-logs) and review-related websites.

2. *Opinion mining*. Opinion mining concerns analyzing the opinion a text expresses. Motivated by real-world applications researchers have considered a wide range of problems in this area (Pang & Lee, 2008). Esuli & Sebastiani (2006) have organized these problems into three categories:

1. Determining *subjectivity*: the problem of determining whether a given text passage has a factual nature or expresses an opinion.
2. Determining *orientation* (also called *polarity*): the problem of determining whether a given subjective text expresses a positive or negative opinion.
3. Determining the *strength of orientation*: for example, weakly positive or strongly negative.

A closely related task is extracting information on *why* the topic or product in the text is considered positive or negative (Pang & Lee, 2008). Other research problems include automatically determining the political colour of a text, for example, liberal or conservative (Mullen & Malouf, 2006). The three categories identified by Esuli & Sebastiani (2006) account for the majority of the research in opinion mining. In this paper we evaluate algorithms for the first two categories: automatically determining subjectivity, and determining the orientation of the subjective passages.

3. *Topical sentiment analysis*. Here the goal is not only to determine that a passage expresses a certain attitude, but also to determine the object of the attitude. Note that in many cases this is known from other sources than the passage (e.g., it is listed separately in a product or movie review). A common approach is to apply a categorization algorithm to a text and then perform a sentiment analysis (Osman & Yearwood, 2007).

4. *Parliamentary proceedings*. More and more large historical corpora of parliamentary proceedings become available for research. Examples include the British Hansard and the Dutch Parlendo. These are two search engines which provide the digitized parliamentary proceedings in a machine-readable XML format. These corpora contain a wealth of information for historical political analysis but digital sustainability and good access to them remains a research challenge (Marx, Aders, Schuth, 2010).

### 3 Data

We test our algorithms on the verbatim transcripts of the plenary meetings of the Dutch House of Representatives. These are available in a variety of technical formats (PDF, Word, HTML, XML). All of these, except the XML format, are meant for human reading and not for machine processing. In this respect the Dutch data is typical for parliamentary proceedings (Marx & Schuth, 2010).

It is a technical challenge to transform some of these formats into useful machine processible formats, especially for the older scanned and OCRed material (Marx & Schuth, 2010). In order to determine political viewpoints from these texts the following information, easily detected by humans but not by machines, is needed: for each word in the text we need to know whether it was spoken or not. If so, by whom, the role or function of the speaker, when, and in which context. For sentiment analysis, we also need a reliable segmentation of the text into words and paragraphs.

These desiderata were best met when using the HTML version of the proceedings data. These are available from the Dutch parliament directly, one day after each meeting, as a draft version (from [www.tweedekamer.nl](http://www.tweedekamer.nl)). The transcripts were downloaded and automatically transformed into the XML format described in Marx & Schuth (2010). An example is provided in the Appendix.

## 4 Assessing subjectivity and orientation

In this section, we present several sentiment analysis techniques, apply them to our dataset, and systematically evaluate their quality using a manually annotated corpus. We perform the first two classification tasks from the schema by Esuli & Sebastiani (2006). First we determine whether a text is subjective or objective. Second, for the subjective texts, we determine whether they have a positive or negative orientation. Before we start with the classifiers we first must decide on the level of detail on which classification is done and on how to create the gold standard corpus.

### 4.1 Classification level

Before classification can commence, the level at which it will be conducted needs to be chosen. Different levels are used in the literature:

- *Document level* (Yu & Hatzivassiloglou, 2003): whole documents are labeled. For example, a document can have an overall orientation that is classified as positive.
- *Block level* (Osman & Yearwood, 2007): the text is cut into several blocks and each block is labeled independently. This is most often used in unstructured data like blog pages.
- *Paragraph level* (Kamps & Marx, 2001): each paragraph is labeled.
- *Sentence level* (Riloff & Wiebe, 2003; Wilson, Pierce, & Wiebe, 2003; Furuse, Hiroshima, Yamada, & Kataoka, 2007): each sentence is labeled.
- *Word level* (Yu & Hatzivassiloglou, 2003; Kim & Hovy, Automatic Detection of Opinion Bearing Words and Sentences, 2005; McKeown & Hatzivassiloglou, 1997): individual words are labeled.

Classification at document level and word level is unsuitable for identifying political viewpoints in parliamentary proceedings. Document level classification means a whole meeting is treated as an individual entity, and marking it will give no particular views of individual parties or political persons. It is too general to be of value. In contrast, classification at the word level is too detailed, and will not contain enough contextual information to connect sentiment to a particular viewpoint or topic.

Classification at the sentence level also has problems with contextual information since individual sentences will contain references to adjacent sentences and topics. For example, the sentence ‘That is okay.’ contains an opinion, but we do not know what the opinion is about. Arguments containing a viewpoint are often expressed in multiple sentences. This leaves us with the choice between either block level or paragraph level classification. Since a paragraph is considered a natural block, this has been considered most appropriate. The source data was already split into paragraphs (using the p-element in HTML), so no further processing was needed.

### 4.2 Gold standard corpus

Evaluation of opinion retrieval algorithms mostly relies on a comparison with human annotations from the same corpus (Ku, Liang, & Chen; Osman & Yearwood, 2007). To evaluate the performance of the algorithms on the Dutch parliamentary proceedings, a gold standard was developed. We aimed at annotating around a thousand paragraphs. For efficiency reasons, the paragraphs were extracted from as few documents as possible. We randomly chose two meetings (March 5<sup>th</sup> and April 21<sup>st</sup>, 2009) which contained enough (1201) paragraphs. Paragraphs spoken by the chairman were not annotated because the

chairman does not take part in the discussions on political issues, but instead tries to keep the meetings on track.

The first task was to annotate whether a paragraph contains an opinion or not. If there is an opinion present, the paragraph is considered subjective. Otherwise the paragraph is considered objective. Examples of objective and subjective paragraphs are given in the Appendix. Two human annotators were used, both with Dutch as their mother tongue. The paragraphs were printed and split evenly between them. A face-to-face explanation of the intention of the research and their task of annotating the paragraphs was provided. They annotated each paragraph as subjective or objective. This was judged by reviewing each individual paragraph against a definition of subjectivity. The definition of subjectivity that was used is, based on the literature (Kim & Hovy, 2004; Riloff & Wiebe, 2003; Wiebe & Riloff, 2005; Wiebe, Bruce, & O'Hara, 1999; Banea, Mihalcea, & Wiebe; Smith, 1999), the *Compact Oxford English Dictionary* definition of *opinion*, and the Dutch *Van Dale online dictionary's* definition of *mening*). Our definition is:

If the primary intention of a piece of text is an objective presentation of material that is factual to the reporter, and does not contain a judgment or emotion, the text is objective. Otherwise the text is subjective.

The second task was to annotate the semantic orientation of each subjective paragraph. As mentioned, the orientation of a text is whether it expresses a positive or negative opinion. The same two annotators were used. This time, however, instead of splitting the paragraphs evenly between them, the two annotators individually marked all of the subjective paragraphs. Discontinuities between the annotators were afterwards resolved via mutual consultation. To have two judgments meant that the inter-annotator agreement could be monitored (see below).

In the literature, a clear definition of positive and negative orientation is hard to find. Most of the time multiple human annotators are used to judge a corpus based on their intuition or common sense (Jijkoun & Hofmann, 2009; Furuse, Hiroshima, Yamada, & Kataoka, 2007). Based on research by Osgood, Suci, & Tannenbaum (1957) the semantic orientation on which we wish to classify the paragraphs is the evaluative factor: good/bad. They proved that this factor is the most significant influence on variation in data. A definition of orientation based on this research can be found in Turney (2001): "a phrase has a positive semantic orientation when it has good associations and a negative semantic orientation when it has bad associations". Turney & Littman (2003) also distinguish between positive evaluation (e.g., praise) and negative evaluation (e.g., criticism) respectively. From these sources, the following definition to classify this binary orientation was formulated, leaning heavily on Osgood, Suci, & Tannenbaum (1957). The annotators were instructed to use this definition in their task:

A text has a positive orientation when it has good associations, or contains a positive evaluation (e.g., praise). The text has a negative orientation when it has bad associations or contains negative evaluations (e.g., criticism).

*Corpus statistics.* In the transcripts of the meetings of March 5<sup>th</sup> and April 21<sup>st</sup>, 2009, a total of 1201 paragraphs were annotated, of which 590 (49.1%) were annotated as subjective. Out of these 590 subjective paragraphs, 251 (42.5%) were annotated as positive, and 339 (57.5%) as negative. Because two annotators were used, inter-annotator agreement could be calculated. The overall agreement is 71.4%. There was hardly any difference in agreement between the positive and the negative paragraphs (175/251=69.5% and 246/339=72.4%, respectively). Cohen's  $\kappa$  is 0.423.

*Conclusions. Because of the use of definitions, the annotation tasks were easy to explain to the annotators. Also, because strict definitions were used, the annotators did not need to have specific domain knowledge. According to some, an analytic definition of opinion is impossible (Kim & Hovy, 2004). Still, even with the strict instructions, inter-annotator agreement was low: overall agreement on semantic orientation between the two annotators was 71.4% and  $\kappa = 0.423$ . These rather low values are however common for sentiment analysis tasks: e.g., (Kim & Hovy 2005) classified 174 sentences by three annotators and found a pair-wise agreement of 73% and a kappa value of 0.49.*

### 4.3 Automatically determining subjectivity

We now describe and evaluate a number of algorithms for determining subjectivity of texts. Technically, these algorithms are binary classifiers: for each input they decide whether it is subjective or not. Algorithms for such a task fall into two categories: based on handcrafted rules, or (machine) learned from examples. Algorithms in the first category make use of so called subjectivity lexicons. We will evaluate two algorithms based on lexicons and three based on machine learning.

Algorithms based on subjectivity lexicons.

*In this approach, the focus is on the number of occurrences of each term. The exact ordering of the terms in a text is not important (Manning, Raghavan, & Schütze, 2008). Most often the individual words are given a certain subjectivity score based on a set of opinion words (the subjectivity lexicon) (Ding & Liu, 2007). The models then present a way to calculate the subjectivity of the whole text based on the individual collection and frequency of these words. We discuss two models from Kim & Hovy (2005) which implement this idea.*

Model 1 counts the total valence score of all words in the paragraph. The basis of this model is that paragraphs dominated by words considered to be subjective tend to be opinion bearing. Individual words in the paragraph are extracted and given a score of 0, 1 or 2, in which a score of 2 is considered to be very subjective and a score of 0 not subjective. A Dutch sentiment wordlist developed by Jijkoun & Hofmann (2009) was used to rate the words. Words not present in the wordlist are considered to be not subjective and have been given a score of 0. A cut-off threshold had to be selected in order to determine when a paragraph is judged to be subjective or objective. Experimentation has been conducted with threshold values between 0 and 20.

Model 2 checks the presence of a single strong valence word. The assumption underlying this model is that the presence of one strong valence word is enough to indicate subjectivity. The Jijkoun & Hofmann sentiment lexicon (2009) is used and a cut-off threshold is set to determine at which score a paragraph is considered to be subjective. Because the wordlist by Jijkoun & Hofmann (2009) contains scores of 0, 1 and 2, where 0 indicates neutrality, the performance of the algorithm is evaluated on cut-off thresholds of 1 and 2.

#### *Machine-learning algorithms*

Machine-learning algorithms differ from models based on subjectivity lexicons in that they automatically train themselves to classify the data. The methods we use are called supervised methods because a labeled data set is needed to train the classifier. For this we use our gold standard. The following machine-learning algorithms are used (Manning, Raghavan & Schütze 2008):

- NaiveBayes;
- IBk nearest-neighbour, with  $k=1$ ;

- Support Vector Machine (SVM) SMO;
- ZeroR (as the baseline).

The toolkit Weka 3.6.1 is used to train and evaluate the machine-learning classifiers.

### Results

The performances of the algorithms are based on accuracy, precision, recall and F-measure (Manning, Raghavan & Schütze 2008). The machine learning algorithms are evaluated using ten fold cross-validation.

As described above, Model 1 uses a cut-off threshold above which the text is classified as subjective. The performance of the model at different cut-off thresholds can be found in Table 1. The highest scores are in bold font. Because the aim of the subjectivity classification is to retrieve paragraphs that are subjective, we are interested mostly in the results of the TRUE-class. (The TRUE class consists of the set of subjective paragraphs). A higher threshold will lower recall, and will increase precision. This is as expected, as a higher cut-off threshold will include less subjective markers (Kim & Hovy, 2005). The table should be interpreted as follows: A threshold of  $> n$  means that a paragraph is classified as subjective if the sum of the valence scores in the paragraph is larger than  $n$ . Thus the threshold " $>0$ " means that a paragraph is subjective if it contains at least one word with subjectivity score 1. We see that with threshold " $>1$ ", we find 93% of all subjective paragraphs (recall=.929) and that 55% of those classified as subjective are indeed subjective (precision=0.550).

Model number 2, based on Kim & Hovy (2005), also uses a cut-off threshold. Here the thresholds have a different meaning: threshold 1 means that the paragraph contains at least one subjective word (thus with score 1 or 2); threshold 2 that it contains at least one word with score 2 (a highly subjective word). The results are in Table 2. Threshold 1 has the best tradeoff between precision and recall (F=.677). Note that Model 2 with threshold 1 is exactly the same as model 1 with threshold  $> 1$ : both classify a paragraph as subjective if it contains at least one subjective word.

The best results of the two models and the results of the machine-learning algorithms are shown in Table 3. In reaching these results, experiments were conducted for smoothing them out, for example, all words in the paragraphs were converted to lowercase. These experiments did not improve results significantly. In fact, converting the paragraphs to lowercase even deteriorated results.

Table 1: Results of cut-off threshold values using model 1 based on Kim & Hovy (2005).

Threshold	Results on TRUE class		
	Precision	Recall	F-measure
Baseline (all subjective)	0.491	1.0	.659
$> 0$	0.521	<b>0.966</b>	0.677
$> 1$	0.550	0.929	<b>0.691</b>
$> 2$	0.570	0.854	0.684
$> 3$	0.591	0.775	0.671
$> 4$	0.605	0.686	0.643
$> 5$	0.636	0.615	0.625
$> 6$	0.653	0.546	0.595
$> 7$	0.671	0.478	0.558
$> 8$	0.671	0.395	0.497
$> 9$	0.684	0.337	0.452
$> 10$	0.689	0.275	0.393
$> 11$	0.695	0.232	0.348
$> 12$	0.688	0.186	0.293

> 13	0.705	0.146	0.242
> 20	<b>0.793</b>	0.039	0.074

Table 2: Results of cut-off threshold values using model 2 based on Kim & Hovy (2005)

Threshold	Results on TRUE class		
	Precision	Recall	F-measure
1	0.521	<b>0.966</b>	<b>0.677</b>
2	<b>0.596</b>	0.666	0.628

Table 3: Results of all approaches on classifying subjectivity (using optimal threshold results at K&H models for weighted results and TRUE class results)

Model	Results on TRUE class		
	Precision	Recall	F-measure
K&H model 1 (threshold >1)	0.550	<b>0.929</b>	<b>0.691</b>
K&H model 2 (threshold 1)	0.521	0.966	0.677
NaiveBayes	0.607	0.802	<b>0.691</b>
IBk	0.563	0.593	0.578
SMO	<b>0.638</b>	0.610	0.624

### Conclusions

The performance of Model 1 on the TRUE class is amongst the highest with an F-measure of 0.691: it finds almost all subjective paragraphs and classifies just over half of them correctly. How good are these results? For that we compare our scores to those obtained by Kim & Hovy (2005). However they report F measures for the weighted results of both classes. We did not report on these measures in the above Tables, but we have calculated them in. The F-measure of Model 1 on our data is at its peak at 0.545 (threshold > 6). Our implementation of Model 1 performs better on our data than the original model performed on TREC 2003 data that only achieved a F-measure of 0.425. The implementation of Model 2 also performed better on our data than the original model on TREC 2003 data, achieving a F-measure of 0.534 (threshold 2) as opposed to 0.514. Thus we can conclude that the results of the methods based on subjectivity lexicons are very promising, as they perform relatively well on political data.

From the machine-learning algorithms, NaiveBayes performs best overall with a weighted F-measure of 0.640 and an F-measure on the TRUE class of 0.691. If we select ZeroR, which predicts a class based on the mode (thus in our case: it classifies all paragraphs as objective), as baseline, NaiveBayes performs significantly better than ZeroR (*one tailed test, confidence level 0.99*). The SVM algorithm SMO also produces decent results significantly better than ZeroR (*one tailed test, confidence level 0.99*). The weighted F-measure of 0.638 comes in range of the NaiveBayes' weighted F-measure of 0.640. On classifying the TRUE class, however, NaiveBayes would still be the preferred algorithm of choice with a F-measure of 0.691 as opposed to SMO's F-measure of 0.624.

#### 4.4 Automatically determining semantic orientation

We now evaluate algorithms which automatically determine semantic orientation. Like the algorithms determining subjectivity, these algorithms are classifiers. In line with the literature, we built binary classifiers which classify subjective paragraphs as positive or negative. Again there are lexicon based approaches and machine learning algorithms.

##### *Algorithms based on subjectivity lexicons*

The algorithm is based on the model by Edens, Liem, Mensink, Weve & Zande (2006) and uses the wordlist by Jijkoun & Hofmann (2009). The algorithm classifies all words as positive, negative or neutral. The scores +2 and +1 are considered positive, and -1 or -2 are considered negative. The algorithm also takes into account that two adjacent polar words of the same orientation influence each other. A factor is calculated based on the distance between the two polar words, with a maximum distance of 10. The score of the original polar word is then multiplied by this factor. The following equation is used to calculate the new wordscore:

$$wordscore = wordscore \cdot \left(1 + \frac{10/distance}{10}\right)$$

After all the wordscores in the paragraph have been calculated, they are added up. If the final score is above 0, the paragraph is classified as positive, otherwise the paragraph is negative.

The model based on Chesley, Vincent, Xu, & Srihari (2006) combines a subjectivity lexicon and machine learning. The expectation of this model is that the distribution of positive and negative adjectives, and positive and negative verb classes, shows regularities. Furthermore, it assumes that the orientation of adjectives can be described by the majority orientation class of their synonyms. We have implemented this model as follows:

First, a part of speech (POS) tagger (TreeTagger 3.2) is used to identify all verbs and adjectives. Next, for all adjectives, the synonyms are scraped from the website [www.synonyms.net](http://www.synonyms.net). In the original implementation by Chesley, Vincent, Xu, & Srihari (2006), Wikipedia's dictionary is used because of its coarse-grained content. We used [www.synonyms.net](http://www.synonyms.net) instead because the Dutch version of Wikitionary is not sufficiently developed yet. All collected synonyms are matched against a wordlist of positive and negative adjectives. The wordlist is created by merging the adjectives of Jijkoun & Hofmann (2009) and the negative and positive adjectives collected by Kamps & Marx (2001). The majority class of the synonyms has been assigned to the adjective. The verbs are assigned to a positive or negative class based on the lexicon by Jijkoun & Hofmann (2009). As output, the model provides a list of information on each paragraph consisting of:

1. Number of positive adjectives
2. Number of negative adjectives
3. Number of positive verbs
4. Number of negative verbs

To this list, the gold standard classification on orientation belonging to the paragraph is added. Finally, using the information gathered on the paragraphs, Chesley, Vincent, Xu, & Srihari (2006) used a SVM algorithm to classify the paragraphs. They opt for the use of a SVM algorithm because they believe it to be robust for sentiment classification and handling noisy data (Mishne, 2005). We use Weka's SVM algorithm SMO, but experiment with NaiveBayes, IB1, and ZeroR as well (see Table 4). NaiveBayes gave the best results.

Table 4: Results of classification on orientation using the output of the model based on Chesley, Vincent, Xu, & Srihari (2006)

Classifier	Precision	Recall	F-measure
------------	-----------	--------	-----------

SMO	0.567	0.581	0.560
NaiveBayes	<b>0.597</b>	<b>0.602</b>	<b>0.599</b>
IB1	0.553	0.558	0.554
ZeroR	0.330	0.575	0.419

### Machine-learning algorithms

Again, four machine-learning algorithms are selected to represent this category.

- NaiveBayes
- IB1 nearest-neighbour
- Support Vector Machine (SVM) SMO
- ZeroR

Again, the Weka toolkit is used to evaluate these algorithms.

### Results

The machine-learning algorithms are evaluated using ten fold cross-validation. Similarly to the algorithms determining subjectivity, all algorithms are evaluated on precision, recall and F-measure. Because the four features used by the model based on Chesley, Vincent, Xu, & Srihari (2006) are numeric, discretization of the data could be conducted to improve results. Experiments have been conducted with different bin sizes for each classifier. The optimal results can be found in Table 4. The following parameters were used:

- SMO: Discretization with Weka's option findNumbins. This option lets Weka choose an appropriate amount of bins.
- IB1: no discretization.
- NaiveBayes: no discretization.

NaiveBayes produces the best results on all measures: almost 60% of all classifications is correct at a recall value of .602. It is difficult to compare these results to the results found by Chesley, Vincent, Xu, & Srihari (2006) since they classify on document level of a blog post instead of on paragraph level. The results nevertheless, allow us to conclude that the SMO classifier is performing well on the data. The results support the claim of Chesley, Vincent, Xu, & Srihari (2006) and Mishne (2005) that it is a robust classifier for sentiment classification.

A comparison of all algorithms used to classify semantic orientation can be found in Table 5. Some smoothing experiments were conducted, but again did not improve results. In contrast to subjectivity results, the SMO machine learning classifier scores best on all fronts regarding orientation classification. It is followed by NaiveBayes. Both perform significantly better than the other models used (*one tailed test, confidence level 0.99*). The combination of collecting paragraph statistics and using a machine learning algorithm gives promising results with a F-measure of 0.599. If more characteristics are collected, performance may increase.

Table 5: Results of classifications by semantic orientation

Model	Precision	Recall	F-measure
model based on Edens, Liem, Mensink, Weve, & Zande (2006)	0.369	0.517	0.419
model based on Chesley, Vincent, Xu, & Srihari (2006) with NaiveBayes	0.597	0.602	0.599
IBk	0.601	0.561	0.556
NaiveBayes	0.652	0.651	0.652
SMO (SVM)	<b>0.677</b>	<b>0.676</b>	<b>0.677</b>
ZeroR	0.330	0.575	0.419

## 5 Conclusions

The aim of this chapter was to evaluate the appropriateness of sentiment analysis techniques which are developed for blogs and product reviews, for the analysis of political texts. We first summarize our technical results and then return to their impact on the main research question.

We studied six algorithms to automatically detect opinion-bearing paragraphs. Machine-learning and lexicon based algorithms scored about equally well. NaiveBayes performed best with a weighted F-measure of 0.640 and a F-measure on the TRUE class of 0.691. Model 1, based on subjectivity lexicons, achieved exactly the same F-measure. Next, six algorithms were studied which automatically detect the orientation of the subjective paragraphs. The algorithms classified paragraphs as either positive or negative. Machine-learning algorithms again dominated the results. NaiveBayes reached a F-measure of 0.652, but the SVM implementation in Weka called SMO performed best with a F-measure of 0.677. Both performed significantly better than the other algorithms. These results support the claim that SVM provide a solid method for sentiment classification (Chesley, Vincent, Xu, & Srihari, 2006; Mishne, 2005). It can furthermore be concluded that a model collecting paragraph characteristics and then classifying the paragraphs using machine-learning algorithms provides promising results.

Considering the performances of the classification algorithms, we conclude that results are approximately in line with results found in the literature. An F-measure approaching 0.7 is a common achievement, and can therefore be considered to be a respectable result. In other words, opinion-mining techniques are suitable to automatically retrieve subjective paragraphs from texts and to annotate their orientation. This shows that today's opinion-mining techniques can be successfully applied to Dutch, political, semi-structured transcripts.

With both techniques scoring about equally well, we advise on using either lexicon based or machine-learning for political texts based on other grounds. It seems that a machine learning approach might be preferable. First, using Crowdsourcing platforms like Amazon Mechanical Turk, it becomes easy, fast, and inexpensive to create a large number of training examples. Even though the task is difficult and we cannot expect high inter-annotator agreement, this option can work fine. In usual experiments, each task is performed by five annotators. Only examples with a strong majority agreement (e.g., 4 out of 5 agree) could be used. Second, politicians are creative language users, and political language as a result changes fast. It is therefore advisable to train a learning algorithm on examples, rather than trying to capture the political language in a lexicon, especially since orientation in political data is topic-dependent. One note of caution is called for though. In case of a lack of resources, as in this study, the machine-learning algorithms are trained and tested on the same set. This might have the effect of overtraining the algorithms.

How can these results be used to retrieve political viewpoints or party positions? For retrieving a viewpoint we also need to know about which topic an opinion was given. This can be done by combining a topical paragraph classifier with the here developed sentiment classifier. Then there are still several ways to distill a party position on a topic based on a collection of paragraphs about that topic spoken by many members of that party. Thus a separate evaluation seems needed to answer the question.

The chapter by Hirst et al. (this volume) finds that lexical methods (as we use here) do not retrieve party positions but rather government versus opposition positions. Whether this also holds for opinionated paragraphs is a matter for further research. Here we can only

give a first impression. We have analyzed our two days of annotated debates with respect to this question and obtained mixed results<sup>1</sup> (see Table 6). The debates of April 21, 2009 show an almost perfect dichotomy between speakers from the government and parties in the government (ChristenUnie, CDA and PvdA), and the other parties. The only exception is “Verdonk”, with just 6 paragraphs. All opposition parties have more negative than positive paragraphs, while it is exactly the other way around in the other group. The other day does not have this clear division between government and opposition. On this rather gloomy day (on which even the government has almost twice as many negative paragraphs than positive), only the Christian parties (CDA, ChristenUnie and SGP) and the Greens (GroenLinks) are more positive in general.

Table 6: Results of classifications by semantic orientation

Thursday March 5th, 2009				Tuesday April 21st, 2009			
Party	POS/NEG	POS	NEG	Party	POS/NEG	POS	NEG
ChristenUnie	1.3	4	3	ChristenUnie	3.7	11	3
CDA	1.1	27	24	Verdonk	3.0	3	1
GroenLinks	1.0	14	14	CDA	1.8	25	14
SGP	1.0	6	6	PvdA	1.6	11	7
PvdA	0.7	15	21	Government	1.2	20	17
D66	0.7	10	15	SGP	1.0	5	5
Government	0.6	33	59	PVV	0.9	12	13
VVD	0.5	16	34	GroenLinks	0.8	7	9
PVV	0.4	4	11	VVD	0.8	10	13
SP	0.2	10	41	SP	0.3	8	28
Verdonk	0.0	0	0	D66	0.0	0	0
PvdD	0.0	0	0	PvdD	0.0	0	1

*EXPLANATION: We count the number of positive and negative paragraphs spoken by members of that party for each day, for each party. The second attribute is the ratio between the numbers of positive and negative paragraphs. If the ratio is higher than 1, the party is overall positive that day, below 1 indicates a more gloomy day.*

Another factor that heavily influences the findings is the style of a debate. In the Dutch House of Commons, interruptions of speeches are made frequently and are recorded in the proceedings together with answers to interruptions. This may cause that governing parties also use more negative terms (e.g., in their answers to negatively phrased interruptions). An interesting dataset to test the findings of Hirst et al. are the Danish parliamentary proceedings. Denmark has a minority cabinet since the early nineties. It could be interesting to see how the dynamic between government and parliament in this situation is reflected in the language of attack and defense.

## References

<sup>1</sup> The editors of this volume made the following notes on these two days. April 21, the day that you find a perfect division between government and parliament, is a Tuesday. On Tuesdays the Dutch parliament has Question Time. This might explain the division between government and the government-parties on the one hand and parliament, more specifically the opposition parties, on the other hand. The opposition parties attack the government as much as possible and the government praises itself and is praised by the government parties (or at least addresses it more positively). March 5th is a Thursday with more elaborate and detailed debates in which parties (apparently) take less clear positions. Please note that Hirst has consistently reported results for Oral Question Period and Debates separately. The results or OQP are usually higher than for Debates (and never lower).

- Bailey, J., & Fekete, A. Eds. *ACM International Conference Proceeding Series*. 242, pp. 133-139. Darlinghurst, Australia: Australian Computer Society.
- Banea, C., Mihalcea, R., & Wiebe, J. A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. *LREC 2008*.
- Boiy, E., Hens, P., Deschacht, K., & Moens, M.-F. (2007). Automatic sentiment analysis of on-line text. *Proceedings of the 11th International Conference on Electronic Publishing*, (pp. 349-360). Vienna, Austria.
- Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). *Using Verbs and Adjectives to Automatically Classify Blog Sentiment*. AAAI Spring Symposium Technical Report SS-06-03.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37-46.
- Compact Oxford English Dictionary: opinion*. (n.d.). (O. U. Press, Producer) Retrieved 06 05, 2009 from Compact Oxford English Dictionary: [http://www.askoxford.com:80/concise\\_oed/opinion](http://www.askoxford.com:80/concise_oed/opinion)
- Ding, X., & Liu, B. (2007). The utility of linguistic rules in opinion mining. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 811-812). Amsterdam, The Netherlands: ACM, New York, NY.
- Edens, J., Liem, M., Mensink, T., Weve, R., & Zande, L. v. (2006). *Measuring Politics*. University of Amsterdam, Amsterdam. *Eindhoven Corpus*. (n.d.). From Instituut voor Nederlandse Lexicologie - Eindhoven Corpus: [http://www.inl.nl/index.php?option=com\\_content&task=view&id=350&Itemid=579](http://www.inl.nl/index.php?option=com_content&task=view&id=350&Itemid=579)
- Esuli, A., & Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. *Proceedings of the Eleventh Conference on European Chapter of the Association for Computational Linguistics* (pp. 193-200). Trento, Italy: European Chapter Meeting of the ACL. Association for Computational Linguistics.
- Fangzhong, S., & Markert, K. (2008). From Words to Senses: a Case Study in Subjectivity Recognition. *Proc. of Coling 2008*. Manchester, UK.
- Furuse, O., Hiroshima, N., Yamada, S., & Kataoka, R. (2007). Opinion sentence search engine on open-domain blog. *Proc. of 20th Int. Joint Conf. of Artificial Intelligence (IJCAI2007)*.
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 22-25). Seattle, WA, USA.
- Jijkoun, V., & Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. *2th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*.
- Kamps, J., & Marx, M. (2001). Words with Attitude. *1st International WordNet Conference*, (pp. 332-341).
- Kamps, J., Marx, M., Mokken, R., & Rijke, M. d. (2004). Using WordNet to measure semantic orientations of adjectives. *Proceedings LREC*.
- Kim, S.-M., & Hovy, E. (2004). Determining the Sentiment of Opinions. *Proceedings of COLING-04*, (pp. 1367-1373). Geneva, Switzerland.
- Kim, S.-M., & Hovy, E. H. (2005). Automatic Detection of Opinion Bearing Words and Sentences. *Second International Joint Conference on Natural Language Processing*.
- Ku, L., Liang, Y., & Chen, H. Tagging heterogeneous evaluation corpora for opinionated tasks. *LREC 2006*.
- Kushal, D., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, (pp. 20-24). Budapest.
- Liu, B. (2007). *Web Data Mining: Exploring hyperlinks, contents and usage data*. Springer.
- Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Marx, M., Aders N. & Schuth, A. (2010). Digital sustainable publication of legacy parliamentary proceedings. *In Proceedings dg.o 2010*.
- Marx, M. & Schuth, A. (2010). DutchParl. A corpus of parliamentary proceedings in Dutch. *In Proceedings LREC 2010*. (pp. 3670-3677).
- McKeown, K., & Hatzivassiloglou, V. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual meeting of ACL*.
- Mishne, G. (2005). Experiments with mood classification in blog posts. *st Workshop on Stylistic Analysis Of Text For Information Access*.
- Morgeson, F. P., & Nahrgang, J. D. (2008). Same as It Ever Was: Recognizing Stability in the BusinessWeek Rankings. *Academy of Management Learning & Education*, 7 (1), 26-41.
- Mullen, T., & Malouf, R. (2006). A preliminary investigation into sentiment analysis of informal political discourse. *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, (pp. 159-162).
- Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.
- Osman, D., & Yearwood, J. (2007). Opinion search in web logs. *Proceedings of the Eighteenth Conference on Australasian Database*, 63. Ballarat, Victoria, Australia.
- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2 (1-2), 1-135.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing* (pp. 79-86). Association for Computational Linguistics.
- Rahm, E., & Do, H. (2000). Data Cleansing: Problems and Current Approaches. *IEEE Data Engineering Bulletin*.
- Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, (pp. 105-112).
- Smith, H. F. (1999). Subjectivity and Objectivity in Analytic Listening. *Journal of the American Psychoanalytic Association*, 47 (2), 465-484.
- TreeTagger 3.2*. (n.d.). From TreeTagger 3.2: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
- Turney, P. (2001). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association For Computational Linguistics* (pp. 417-424). Philadelphia, Pennsylvania: Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ.
- Turney, P., & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21 (4), 315-346.
- Tweede Kamer: plenaire vergaderingen*. (n.d.). Retrieved 06 03, 2009 from Tweede Kamer der Staten Generaal: [http://www.tweedekamer.nl/vergaderingen/plenaire\\_vergaderingen/index.jsp](http://www.tweedekamer.nl/vergaderingen/plenaire_vergaderingen/index.jsp)

- Van Dale online dictionary: mening* . (n.d.). (V. Dale, Producer) Retrieved 06 05, 2009 from Mening: <http://www.vandale.nl/vandale/opzoeken/woordenboek/?zoekwoord=mening>
- WekaWiki: Primer*. (n.d.). Retrieved 06 10, 2009 from Weka---Machine Learning Software in Java: <http://weka.wiki.sourceforge.net/Primer>
- WekaWiki: Text categorization with Weka*. (n.d.). Retrieved 06 08, 2009 from Weka---Machine Learning Software in Java: <http://weka.wiki.sourceforge.net/Text+categorization+with+Weka>
- Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246--253). College Park, Maryland: Association for Computational Linguistics.
- Wiebe, J., & Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Computational Linguistics and Intelligent Text Processing* (Vol. 3406/2005, pp. 486-497). Heidelberg: Springer Berlin.
- Wilson, T., Pierce, D., & Wiebe, J. (2003). Identifying opinionated sentences. *Proceedings of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology: Demonstrations. 4*, pp. 33-34. Edmonton, Canada: North American Chapter Of The Association For Computational Linguistics. Association for Computational Linguistics.
- Witten, I., & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.
- Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, (pp. 129-136).

## Appendix

```

<spreker pagina="" anker="252" naam="Van Raak" partij="SP" soort="Kamerlid"
  geslacht="man">
  <p gs:subjective="false">Voorzitter. 11 commissarissen van de Koningin hebben 35
    duurbetaalde bijbanen. Blijkbaar hebben ze niks anders te doen. Ik heb het heel
    druk, net als de minister. De minister heeft geen duurbetaalde bijbanen, maar die
    commissarissen wel. Waarom eigenlijk? Waarom hebben zij het niet net zo druk als
    deze minister? Wij moeten ze aan het werk zetten.</p>
  <p gs:subjective="true" gs:orientation="negative">Blijkbaar is een aantal deelnemers
    van het ABP, terugblikkend, toch niet zo tevreden over de heer Borghouts. Dat kan.
    Zij vragen nu advies aan Nout Wellink van DNB. Ik steun dat; ik hoop dat de minister
    haar politiemensen, die ook ernstige bezwaren tegen deze benoeming hebben, steunt.</p>
  <p gs:subjective="true" gs:orientation="negative">Ik heb de minister gevraagd om
    een keuze te maken. Dat doet zij echter niet. Als zij geen keuze maakt, maakt
    zij toch een keuze. Ik heb gezegd: als wij de commissarissen serieus nemen, dan
    moeten wij ze ook aan het werk zetten. Dat betekent dus dat zij geen bijbanen kunnen
    hebben. Ik neem de minister serieus en zij heeft ook geen bijbanen. Zeggen wij
    daarentegen: ja, die commissarissen van de Koningin kunnen wel al die bijbanen
    hebben, dan is het CdK-schap geen serieuze baan.</p>
</spreker>
<spreker pagina="" anker="257" naam="Verbeet" partij="" soort="Voorzitter"
  geslacht="vrouw">
  <p>Dank u wel.</p>
</spreker>
<spreker pagina="" anker="260" naam="Van Raak" partij="SP" soort="Kamerlid"
  geslacht="man">
  <p gs:subjective="true" gs:orientation="negative">Ik vraag dus de minister om een eind
    te maken aan al die duurbetaalde bijbanen van commissarissen van de Koningin. Dat is
    namelijk de enige manier om een serieuze baan van het CdK-schap van te maken.</p>
</spreker>
<spreker pagina="" anker="263" naam="Ter Horst" partij="" soort="Minister" geslacht="">
  <p gs:subjective="false">Voorzitter. Ik neem Kamerleden ook buitengewoon serieus. Zij
    mogen ook bijbanen hebben. Ik weet dus niet of dat een goede redenering is. De
    redenering zou moeten zijn dat het CdK-schap een serieuze baan is die serieus
    wordt betaald, maar dat het nuttig kan zijn -- ook voor de provincie -- dat een
    commissaris van de Koningin bijbanen heeft. Provinciale staten oordelen over het
    nut, een mogelijke belangentegenstelling en een cumulatie van bijbanen.</p>
  <p gs:subjective="true" gs:orientation="positive">Voor een groot deel ben ik het eens
    met de heer Van Raak over het financiële aspect. Daarom heeft het kabinet besloten
    om een regeling te treffen voor de neveninkomsten van de commissarissen van de
    Koningin.</p>
</spreker>

```

Figure 1: Example of the data format in XML annotated with the gold standard.