

Opinion mining in Dutch Hansards

Steven Grijzenhout and Valentin Jijkoun and Maarten Marx¹
sgrijzen@science.uva.nl, jijkoun@science.uva.nl, maartenmarx@uva.nl
ISLA, Informatics Institute, University of Amsterdam

Abstract

The question is addressed if opinion mining techniques can be successfully used to automatically retrieve political viewpoints in Dutch parliamentary publications. Two specific tasks are identified: automatically determining subjectivity in the publications and automatically determining the semantic orientation of the subjective parts. A collection of recent parliamentary publications has been collected and a golden standard annotation is created on both subjectivity and orientation. Following this, models based on subjectivity lexicons and machine learning algorithms are evaluated on automatic classification. Overall results tend to be dominated by machine learning algorithms, but methods based on subjectivity lexicons provide promising results. Based on the results it is concluded that opinion mining techniques can indeed be successfully used to automatically retrieve political viewpoints in Dutch parliamentary publications.

Keywords: opinion mining, subjectivity, semantic orientation, parliamentary publications

1 INTRODUCTION

The Dutch House of Representatives is made up out of 150 Members of Parliament all of which adhere to various political parties. One of the main requirements of democracy is that the public needs to be informed about the points of view of each political party concerning all political issues. There are various services trying to summarize and evaluate these points of view and provide custom advice to the voter during elections. Examples of this kind of services include Stemwijzer.nl, Kieskompas.nl and Verkiezingskijker.nl.

Opinion mining is a recent discipline concerned with automatically determining the opinion a text expresses. This paper presents the findings of research conducted to review if opinion mining techniques can successfully to be used to help explore the points of view in Dutch politics. This way, services concerning political data can be improved. But also broader application can be found in for example review-related websites and business and government intelligence (Pang & Lee, Opinion Mining and Sentiment Analysis, 2008).

In section two a problem definition is formulated. In section three a background of relevant research areas is described. In section four an overview is presented of the data used, and the harvesting process used in collecting it. In section five the process of evaluating opinion identification algorithms is described. In section six the conclusions concerning the main research question are presented.

2 PROBLEM DEFINITION

Dutch parliamentary publications contain transcriptions of spoken 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. Furthermore, they provide an indispensable source of information on the interpretation of Dutch law since they contain the perceived purposes of specific laws as discussed in the House of Representatives.

¹ Maarten Marx acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement FOX, number FP7-ICT-233599.

Besides valuable information the transcripts also contain information that is less relevant, and they are often long and quite boring (Marx, 2009). It would make matters easier and save time if the relevant information could be extracted or annotated in the texts in order to make quick information access possible. Information most valuable to services addressing political information as discussed are the different political positions on various issues. This paper concentrates on these positions, and tries to discover if opinion mining techniques can be helpful in automatically addressing these viewpoints. The research is build around one central research question:

Can opinion mining techniques be successfully used to automatically retrieve political viewpoints in Dutch parliamentary publications?

The approach used to find a clear answer on this question consists out of two steps. First, there must be assessed how reliable and accurate opinion mining techniques are in automatically retrieving said political viewpoints. Therefore the performances of different algorithms are evaluated on accuracy, precision, recall and F-measure. Second, these results will be assessed on a theoretical level whether or not they can be considered reliable and accurate enough for this task. Corresponding to these steps, the following research questions have been formulated:

What is the accuracy, precision, recall and F-measure of different automatic opinion classification algorithms on our corpus?

Can these results be considered to be reliable and accurate enough to be successfully applied?

3 BACKGROUND

This research is part of different research areas: opinion search, opinion mining, topic mining and recent research regarding Dutch parliamentary publications. In this section these research areas are discussed shortly.

3.1 Opinion search

Opinion search as a research area is a relatively new branch of studies. The aim is to enable users to search for opinions on any object (Liu, 2007). However, 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. As most research to date covers opinion search applications in the context of weblogs and review-related websites, these will be discussed briefly.

Customer reviews on the Internet provide a valuable source of information. Reviews available on Amazon.com provide information on all aspects of products, and on a huge number of alternative products. Using these reviews, product-developing companies can gather feedback on their products in a relatively easy and cheap way. Similarly, consumers can learn about products before making a purchase. Furthermore, websites like Amazon.com offer the option to rate products aside from merely writing a review about it. These ratings can be used to evaluate the results of opinion mining techniques by comparing the ratings to the results from opining mining.

Activity regarding opinion search has been heavily concentrated on weblogs. They provide a wealth of opinions and information about recent issues regarding a wide range of topics. Weblogs technically distinguish themselves from other sources by being comprised of unstructured and scattered information. In contrast to this, the Dutch parliamentary

publications are more structured and annotated. This difference in structure could require a different approach. Furthermore, despite the fact that research regarding opinion search in weblogs has been conducted in different languages, including English (Osman & Yearwood, 2007) and Japanese (Furuse, Hiroshima, Yamada, & Kataoka, 2007), to our knowledge, it has not yet been applied to Dutch.

Opinion mining techniques as explored in this paper can be viewed in the larger context of opinion search.

3.2 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, Opinion Mining and Sentiment Analysis, 2008). Esuli & Sebastiani (2006) have organized these problems into three categories:

1. Determining *subjectivity*. The problem of determining whether a given text 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.

Other tasks closely related with these categories can be found as well. For example, extracting information on why the topic or product in the text is considered as being positive or negative (Pang & Lee, Opinion Mining and Sentiment Analysis, 2008). Other problems include automatically determining the political colour of a text, for example liberal or conservative (Mullen & Malouf, 2006).

One could say that the categories identified by Esuli & Sebastiani (2006) account for the majority of the research in opinion mining done so far. Nevertheless it is a broad research area with promising possibilities.

3.3 Topic mining

Topic mining is analyzing topics of texts. It is a research area important to a broad range of applications and industries. For example, news articles can automatically be classified on topics or assigned in correct categories.

Topic mining is a powerful tool when combined with opinion mining techniques. Using both techniques one can automatically determine what a text is about and the author's opinion about the topic. This combination is already used in the literature (Osman & Yearwood, 2007).

Text categorization has traditionally been used to classify documents by topic. In this research, we will also make use of algorithms used in text categorization to classify opinions. Thus topic mining and opinion mining are closely related.

3.4 Dutch parliamentary publications

Analyzing parliamentary publications using automatic techniques has been the subject of other research. Most of this research has been conducted by the ILPS group at the University of Amsterdam. Parliamentary publications are studied, because they possess certain characteristics that are valuable to the Information Retrieval community (Marx, 2009), including the following:

- Large historical corpora. All data from 1814 to present of Dutch parliament will be available in 2010;
- Data integration issues and opportunities;
- Natural corpus for content and structure queries.

4 DATA

This research focuses on opinions in the political arena. The Dutch parliamentary publications of public meetings of the House of Representatives will be used as the data source. For the following reasons it has been decided to limit the data used in this research to the Dutch parliamentary publications:

- Political opinions are often expressed in these public meetings;
- During these meetings recent issues are discussed;
- It is a textual representation of spoken words, in accordance to the process of transcription, and can easily be connected to the actual speaker;
- An XML-annotation scheme for the data is already available (Marx, 2009);
- Harvesting and processing the data contributes to building a useful database for future research.

The transcriptions of the meetings of the House of Representatives are published as HTML on their website within a few days. Using the process of Extraction, Transformation and Load (Rahm & Do, 2000) the HTML is transformed into XML. The ETL-process consists of three steps:

1. *Extraction*. The HTML is scraped from the website and downloaded to the server and stored. These original HTML pages are kept as original;
2. *Transformation*. Using XSLT the HTML of the stored webpage is transformed into XML;
3. *Load*. The XML retrieved from the previous step is stored on the server.

The annotation scheme for meeting notes developed by Marx (2009) was used. The scheme provided an excellent format to process and save the meetings, and was flexible enough to facilitate data enrichment. Every spoken word is marked with (1) the speaker, (2) his party at the time of speaking, (3) his role/function in parliament and (4) the ISO-date. In this research the annotation scheme was enriched with (5) a list of members present at the meeting and (6) the gender of each speaker.

All meetings published since February 10th, 2009 have been collected and transformed. The transformation process is not perfect and thus contains errors. Errors in data enrichment are for example found at adding the appropriate gender to speakers, where 7.8% is unknown. A cake-diagram of these percentages can be found in Figure 1. Regarding the data enrichment of adding the political party the speakers belongs to, 27.9% is unknown.

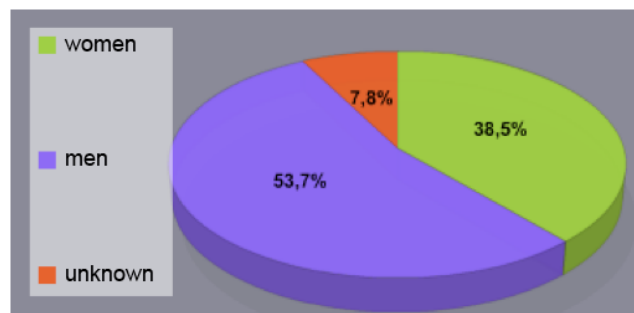


Figure 1: Percentage of gender attribute values on speaker tags

The data enrichment does not affect the selection of data used in this paper. Only paragraph text spoken by members of the House of Representatives has been used. The errors occurring in transformation of these paragraphs do affect this paper. However, the selected data used in this paper is considered sufficiently accurate for the purpose of subjectivity and semantic orientation classification

Furthermore, we can conclude that the data used in this paper differs from data commonly used in opinion mining literature. In this literature, most research is conducted on

weblogs or online reviews. The parliamentary publications in this research differ on the following aspects:

- They are transcriptions, not naturally written texts;
- They are written in Dutch;
- The data is structured to a certain level.

5 OPINION RETRIEVAL

In this section the first sub research question will be addressed: how accurate and precise are the results of automatic opinion classification algorithms on our corpus?

First, the level of detail on which classification is done is being discussed. Second, the process of creating a golden corpus is described. This humanly annotated corpus is used to evaluate the automatic classifications. Third, two types of classifications conducted on the data are discussed. They coincide with two categories presented by Esuli & Sebastiani (2006): subjectivity and orientation.

5.1 Classification level

Before classification can commence, the level at which it will be conducted needs to be chosen. Different levels can be identified in the literature. The following list is ordered from a high level of classification to a low level of classification.

- Document level (Yu & Hatzivassiloglou, 2003): whole documents are labelled. 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 labelled independently. This is most often used in unstructured data like blog pages.
- Paragraph level (Kamps & Marx, 2001): each paragraph is labelled.
- Sentence level (Riloff & Wiebe, 2003; Wilson, Pierce, & Wiebe, 2003; Furuse, Hiroshima, Yamada, & Kataoka, 2007): each sentence is labelled.
- Word level (Yu & Hatzivassiloglou, 2003; Kim & Hovy, Automatic Detection of Opinion Bearing Words and Sentences, 2005; McKeown & Hatzivassiloglou, 1997): individual words are labelled.

Classification at document level and word level is considered to be unsuitable for the aim of identifying political viewpoints in the parliamentary publications. 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 like these 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 as most suitable to the task.

5.2 Golden corpus

Evaluation of opinion retrieval algorithms mostly relies on a comparison of results on the same corpus annotated by humans (Ku, Liang, & Chen; Osman & Yearwood, 2007). To

evaluate the performance of the algorithms on the Dutch parliamentary publications, a golden standard of these meetings has been developed.

Two transcriptions of recent meetings were selected. They contained a total of 1201 paragraphs to be annotated. During annotation, the paragraphs spoken by the chairman were ignored because the chairman does not take part in the discussions of political issues, but instead tries to keep the meetings on track.

The first task was to annotate every paragraph on containing an opinion or not. If there is an opinion present, we consider the paragraph to be subjective. Otherwise the paragraph is considered to be objective. Two human annotators were used. Their native language is Dutch. 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. This definition, shown below, has been formulated based on the use of subjectivity in the literature (Kim & Hovy, Determining the Sentiment of Opinions, 2004; Riloff & Wiebe, 2003; Wiebe & Riloff, 2005; Wiebe, Bruce, & O'Hara, 1999; Banea, Mihalcea, & Wiebe; Smith, 1999), the Oxford Dictionary (Compact Oxford English Dictionary: opinion) and the Dutch dictionary De Grote Van Dale (Van Dale online dictionary: mening).

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 judgement 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. Two annotators were used. Their native language is Dutch. Again, a face-to-face explanation of the intention of the research, and the task of annotating the paragraphs, was provided. 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.

The judgement on subjectivity was once again based on a definition. In the literature a clear definition, if any, 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. Osgood, Suci, & Tannenbaum (1957) 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". In Turney & Littman (2003) they also distinguished between positive evaluation (e.g. praise) respectively negative evaluation (e.g. criticism). From these sources a definition to classify this binary orientation has been formulated, leaning heavily on Osgood, Suci, & Tannenbaum (1957):

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

5.2.1 Results

In the transcripts of the meetings of March 5th and April 21st, 2009, a total of 1201 paragraphs have been annotated. Of them, 590 were annotated as subjective, respectively 49.1%.

Of all 590 subjective paragraphs, 251 were annotated as being positive, and 339 negative. That is 42.5% and 57.5% respectively. Because two annotators were used, inter-annotator agreement can be calculated. The overall agreement is calculated at 71.4%:

$$\text{Overall agreement} = \frac{175 + 246}{590} \approx 71,4\%$$

Cohen's κ is calculated at $\kappa = 0.423$. Cohen's kappa κ takes in account the agreement occurring by chance, and can therefore be considered more robust than simple overall agreement (Cohen, 1960).

5.2.2 Conclusions

Because of the use of definitions, the annotation tasks were easily explainable to other persons. 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 probably impossible (Kim & Hovy, Determining the Sentiment of Opinions, 2004). The use of a definition to review subjectivity could therefore be called into question.

Overall agreement on semantic orientation between the two annotators was 71.4% and $\kappa = 0.423$. A value of κ beneath 0.67 is seen as data providing a dubious basis for evaluative purposes, but the exact values depend on the context of the research (Manning, Raghavan, & Schütze, 2008).

5.3 Automatically determining subjectivity

This section describes the selection, use and results of algorithms to automatically identify paragraphs containing opinions in the data collection. In the literature different categories of classification algorithms can be found. In this research, we use models based on subjectivity lexicons and machine learning algorithms as they are the approaches most commonly used.

5.3.1.1 Models based on subjectivity lexicons

In this approach, the exact ordering of the terms in a text is ignored but it is the number of occurrences of each term that is important (Manning, Raghavan, & Schütze, 2008). Most often the individual words are given a certain subjectivity score based on a set of opinion words previously selected for the purpose (Ding & Liu, 2007). The models then present a way to calculate the subjectivity of the whole paragraph based on the individual collection and frequency of these words.

In this research two models based on subjectivity lexicons have been selected to represent the category. They are chosen because other models are complex and are difficult to implement based on their descriptions in the literature.

Model 1 is based on Kim & Hovy (2005). This model 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 is also based on Kim & Hovy (2005). This model 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 Dutch sentiment lexicon developed by Jijkoun & Hofmann (2009) is once again used. This model too makes use of a cut-off threshold to determine at which score a word is considered to be a word with a high valence. 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.

5.3.1.2 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 labelled data set is needed to train the classifier. For this we will use our golden standard. The following three machine learning algorithms are selected to represent this category:

- NaiveBayes;
- IBk nearest-neighbour, with k=1;
- Support Vector Machine (SVM) SMO;
- ZeroR.

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

5.3.2 Results

The machine learning algorithms are evaluated using a ten fold cross-validation. The performances of all models are based on accuracy, precision, recall and F-measure.

As described above, model number 1 based on Kim & Hovy (2005) 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. They are visualized in Figure 2 and Figure 3. The results include the weighted average results of the TRUE and FALSE classes, and the results on the TRUE class only. 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.

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

Threshold	Weighted results			Results on TRUE class		
	Precision	Recall	F-measure	Precision	Recall	F-measure
0	0.352	0.546	0.414	0.521	0.966	0.677
1	0.375	0.592	0.457	0.550	0.929	0.691
2	0.418	0.612	0.497	0.570	0.854	0.684
3	0.449	0.626	0.522	0.591	0.775	0.671
4	0.474	0.626	0.536	0.605	0.686	0.643
5	0.496	0.638	0.544	0.636	0.615	0.625
6	0.513	0.634	0.545	0.653	0.546	0.595
7	0.530	0.629	0.540	0.671	0.478	0.558
8	0.543	0.608	0.525	0.671	0.395	0.497
9	0.555	0.598	0.512	0.684	0.337	0.452
10	0.564	0.583	0.493	0.689	0.275	0.393
11	0.571	0.573	0.477	0.695	0.232	0.348
12	0.572	0.559	0.456	0.688	0.186	0.293
13	0.584	0.550	0.436	0.705	0.146	0.242
20	0.636	0.523	0.367	0.793	0.039	0.074

[Kim & Hovy, 2005] Model 1 threshold evaluation

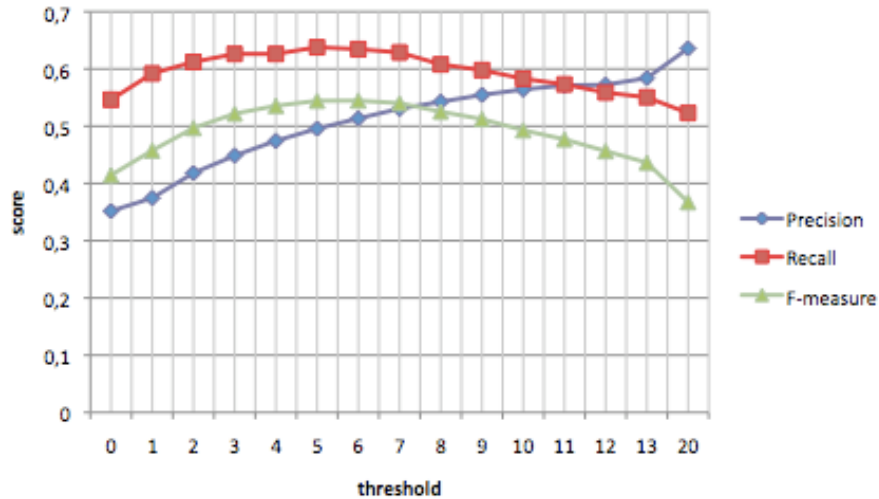


Figure 2: Visualizations of cut-off threshold results as found using model 1 based on Kim & Hovy (2005)

model based on [Kim & Hovy, 2005] Model 1 TRUE-class threshold evaluation

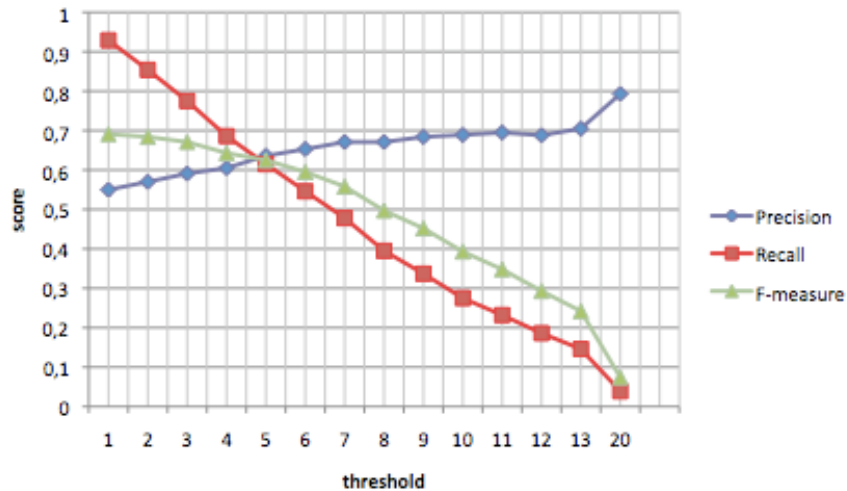


Figure 3: Visualizations of cut-off threshold results as found using model 1 based on Kim & Hovy (2005)

A visualization of the performance results at various cut-off thresholds found by Kim & Hovy (2005) can be found in Figure 4. When compared to the results in Figure 2 and Figure 3, one can see an amount of similarity. 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, Automatic Detection of Opinion Bearing Words and Sentences, 2005).

Original [Kim & Hovy, 2005] Model 1 threshold evaluation

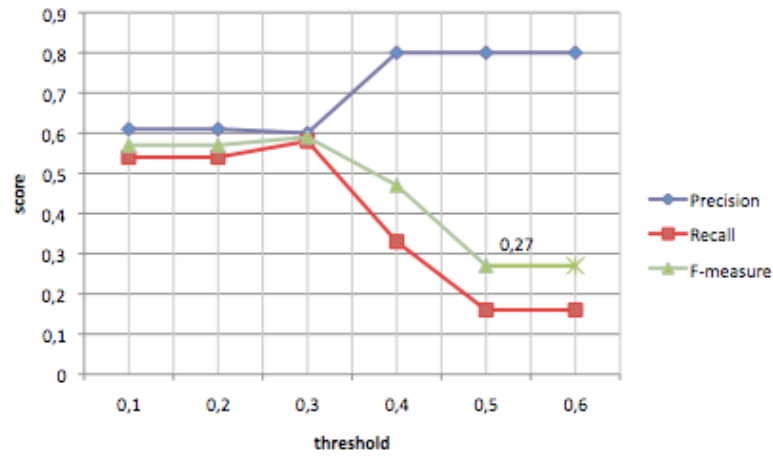


Figure 4: Visualization of cut-off threshold results as found originally by Kim & Hovy (2005)

Model number 2, based on Kim & Hovy (2005), also uses a cut-off threshold. The results can be found in Table 2.

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

Threshold	Weighted results			Results on TRUE class		
	Precision	Recall	F-measure	Precision	Recall	F-measure
1	0.352	0.546	0.414	0.521	0.966	0.677
2	0.478	0.613	0.534	0.596	0.666	0.628

Again, we are mostly interested in the results on classification of the TRUE class. Threshold 1 performs better on the F-measure of the TRUE-class, with a F-measure of 0.677 as opposed to a F-measure of 0.628 of the FALSE-class.

The final results of model 1 and 2, and the results of the machine learning algorithms are shown in Table 3. In reaching these results, experiments with some smoothing have been conducted. An example of this is the converting of all words in the paragraphs to lowercase. These experiments did not improve results significantly. In fact, converting the paragraphs to lowercase even deteriorated results.

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	Weighted results			Results on TRUE class		
	Precision	Recall	F-measure	Precision	Recall	F-measure
K&H model 1	0.513	0.634	0.545	0.550	0.929	0.691
K&H model 2	0.478	0.613	0.534	0.521	0.966	0.677
NaiveBayes	0.666	0.648	0.640	0.607	0.802	0.691
IBk	0.574	0.574	0.574	0.563	0.593	0.578
SMO	0.639	0.639	0.638	0.638	0.610	0.624
ZeroR	0.259	0.509	0.343	0*	0*	0*

* predicts mode, and thus always predicts FALSE, hence TRUE is never classified

5.3.3 Conclusions

The F-measure of model 1 by Kim & Hovy (2005) is at its peak at 0.545. 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. Our 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 as opposed to 0.514. The performance of model 1 on the TRUE class is amongst the highest with a F-measure of 0.691.

One can conclude that the results of the models based on subjectivity lexicons are promising, as they performed relatively well on our target data.

However, NaiveBayes performs best overall with a weighted F-measure of 0.640 and a F-measure on the TRUE class of 0.691. If we select ZeroR, which predicts a class based on the mode, 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.

5.4 Automatically determining semantic orientation

In this section the algorithms and results of automatically determining semantic orientation are discussed. In this research it has been decided to use the binary classification of positive and negative because it is common in the literature. The algorithms can again be categorized as models based on subjectivity lexicons and machine learning algorithms. To represent the categories, a model based on a subjectivity lexicon is selected, four machine learning algorithms are selected, and one algorithm combining both categories is selected.

5.4.1.1 Models based on subjectivity lexicons

An algorithm based on the model by Edens, Liem, Mensink, Weve, & Zande (2006) is implemented to classify all words as positive, negative or neutral. The wordlist by Jijkoun & Hofmann (2009) is used again. Scores of +2 and +1 are considered positive, and -1 or -2 is considered negative. The algorithm also takes into account that two adjacent polar words influence each other. However, this is only implemented for words of the same orientation. 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:

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

After all the wordscores in the paragraph are calculated, they are added up. If the final score is above 0, the paragraph is considered to be positive. If the final score is beneath or equal to 0, the paragraph is considered to be negative.

A combination of a model based on a subjectivity lexicon and machine learning algorithms is used in the model based on Chesley, Vincent, Xu, & Srihari (2006). The reasoning behind this model is that the distribution of positive and negative adjectives, and positive and negative verb classes will show regularities. Furthermore, they believe that the orientation of adjectives can be described by the majority orientation class of their synonyms. In this paper the model is implemented 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. 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 instead of Wikitionary because the latter has not yet been sufficient developed in the Dutch language. All collected synonyms are matched against a wordlist of positive and negative adjectives. The wordlist used 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.

All 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 the following:

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

To this list, the golden standard classification on orientation belonging to the paragraph is added. Finally, using the information gathered about 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 will use Weka's SVM algorithm SMO, but have experimented with NaiveBayes, IB1, and ZeroR as well.

5.4.1.2 *Machine-learning algorithms*

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

- NaiveBayes
- Bk1 nearest-neighbour
- Support Vector Machine (SVM) SMO
- ZeroR

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

5.4.2 *Results*

The machine learning algorithms are evaluated using a ten fold cross-validation. Similarly to the algorithms determining subjectivity, all algorithms are evaluated on precision, recall and F-measure. Because we are interested in paragraphs containing a negative orientation as well as containing a positive orientation, weighted results are most important in this section.

Because the four attributes 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 optimized results can be found in Table 4. The following parameters were used:

- SMO: Discretized with Weka's option `findNumbins`. This option lets Weka choose an appropriate amount of bins.
- IB1: no discretization at all.
- NaiveBayes: also no discretization at all.

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	0.597	0.602	0.599
IB1	0.553	0.558	0.554
ZeroR	0.330	0.575	0.419

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.

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)	0.597	0.602	0.599
IBk	0.601	0.561	0.556
NaiveBayes	0.652	0.651	0.652
SMO (SVM)	0.677	0.676	0.677
ZeroR	0.330	0.575	0.419

5.4.3 Conclusions

The results of classification of paragraph statistics, provided by the model based on Chesley, Vincent, Xu, & Srihari (2006), can be seen in Table 4. NaiveBayes again provides the best results again on all fronts, its precision and recall scoring both the highest. Compared to NaiveBayes, the SMO classifier performs second best with a F-measure of 0.560.

Comparing these results to the results found by Chesley, Vincent, Xu, & Srihari (2006) is difficult since they classify on the document level of a blog post instead of the paragraph level. We could say, however, that the SMO classifier is performing well on our data, and can therefore support the claim of Chesley, Vincent, Xu, & Srihari (2006) and Mishne (2005) that it is a robust classifier for sentiment classification.

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.

6 CONCLUSIONS

We have seen that the ETL-process as described by Rahm & Do (2000) proved to be a reliable way of collecting the official publications available on the website of the Dutch House of Representatives. The scheme to annotate meetings developed by Marx (2009) provided an excellent format to process and save the meetings, and was flexible enough to facilitate data enrichment.

Next, we concluded that the paragraph level was most suitable to our task as it contains enough context information but can also be considered specifically enough. After annotation of the golden corpus, the inter-annotator agreement on orientation was 71.4% and Cohen's kappa was 0.423. These results indicate that the golden standard is a dubious

basis for evaluation purposes (Manning, Raghavan, & Schütze, 2008). This could be an indication of the difficulty of the task.

After this, we have selected six algorithms to automatically detect opinion-bearing paragraphs. The results were mostly dominated by machine-learning algorithms. NaiveBayes performed best with an weighted F-measure of 0.640 and a F-measure on the TRUE class of 0.691. However, we also found that models based on subjectivity lexicons provided promising results as they performed relatively well on our data. Model 1 based on Kim & Hovy (2005) achieved a F-measure on the TRUE class equal to NaiveBayes.

Next, six algorithms were selected to automatically detect the orientation of the subjective paragraphs. This has been done via a binary classification with the classes positive and 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 used. These results support the claim that SVM provide a solid method for sentiment classification (Chesley, Vincent, Xu, & Srihari, 2006; Mishne, 2005). Also, it can be concluded that a model collecting characteristics of a paragraph and then classifying them 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. A F-measure approaching 0.7 is quite a common achievement, and can therefore be considered to be a respectable result. Because of this we can positively answer the question whether the opinion mining techniques can be considered suitable to automatically retrieve subjective paragraphs and annotating their orientation. With this we have shown that today's opinion mining techniques can be successfully applied to Dutch, political, semi-structured transcripts.

References

- 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
- 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
- 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)*.
- Gielissen, T. (2008). *Het ontsluiten van Nederlandse parlementaire publicaties naar Brits voorbeeld*. Bachelor Thesis, University of Amsterdam, FNWI, Amsterdam.
- Grijzenhout, S., Rheenen, E. v., & Marx, M. (2009). *Er wordt wat afgepraat op het Binnenhof, Proefproject Bachelor Thesis Informatiekunde*. Retrieved 06 14, 2009 from http://student.science.uva.nl/~sgrijzen/verslag_warmup/
- 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.
- ILPS. (n.d.). *Homepage*. Retrieved 06 03, 2009 from ILPS information and language processing systems: <http://ilps.science.uva.nl/>
- 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. (2009). Long, often quite boring, notes of meetings. In *Proceedings ESAIR 2009 : Exploiting Semantic Annotations in Information Retrieval*.
- 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
- 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).