

---

## Political sentiment analysis:

### Predicting speaker attitude in the UK House of Commons

Salah, Z. · Coenen, F. · Grossi, D.

**Abstract** In this paper the authors seek to establish the most appropriate mechanism for conducting sentiment analysis with respect to political debates so as to predict their outcome. To this end two alternative approaches are considered, the classification based approach and the lexicon based approach. In the context of the second approach either generic or domain specific lexicons may be adopted, both options are compared with the classification based approach. The comparison between the potential sentiment mining approaches and supporting techniques is conducted by predicting the attitude of individual debaters (speakers) in political debates (using debate transcripts taken from the proceedings of the UK House of Commons). The reported comparison indicates that the attitude of speakers can be effectively predicted using sentiment mining. The authors then go on to consider whether speaker political party affiliation is a better indicator of attitude than the content of the concatenated speeches of individual debaters.

**Keywords** Opinion Mining · Sentiment Analysis · Machine Learning · Information Retrieval

---

Zaher Salah (✉)

Department of Computer Science, University of Liverpool, UK

E-mail: zsalah@liverpool.ac.uk

Frans Coenen

Department of Computer Science, University of Liverpool, UK

E-mail: coenen@liverpool.ac.uk

Davide Grossi

Department of Computer Science, University of Liverpool, UK

E-mail: d.grossi@liverpool.ac.uk

## 1 Introduction

Political analysis, whether this occurs in the form of “official” media (news papers, television reports) or “unofficial” media (blogs, social network sights), is an everyday part of our lives. Consequently the study of political debate is a popular area of sociological and cultural research. For example in (Welch, 1985) a study was undertaken to determine whether US congress women are more liberal than congress men by conducting a study of voting patterns. In (Porter et al, 2005) network analysis techniques were used to determine how the committees and sub-committees of the US House of Representatives were interconnected. The study of political debates is also of interest in terms of how such debates operate, see for example the work of (Rissland, 1999) or (Thomas et al, 2006). In this paper we are interested in techniques to predict the “attitude” of individual speakers, whether they are for or against a motion within the context of political debates, from transcripts of speeches made by individual speakers (debaters). More specifically we are interested in applying sentiment (opinion) mining techniques to predict speaker attitude. This has applications in many contexts such as political campaign management and the practice and theory of argumentation. The focus for the work is the political debates conducted in UK House of Commons.

In general, sentiment (opinion) mining is concerned with the use of data mining techniques to extract positive and negative feelings, opinions, attitudes or emotions, typically embedded within some form of text, concerning some object of interest (Liu, 2012). This object may be a product, a person, some legislation, a movie, or some kind of happening or topic. Sentiment mining is thus directed at the automatic extraction and categorisation of subjective information embedded in various types of textual data as opposed to objective or factual information.

Sentiment mining is typically applied to a “document corpus” comprising either structured or unstructured free text. There are a variety of techniques that can be used for this purpose. One commonly used approach is the classification based approach where a pre-labelled “training” corpus is used to build a classifier that can then be applied to previously unseen texts (Kim and Hovy, 2004; Pang and Lee, 2008) so as to extract the sentiment expressed within these texts. For example in (Dang et al, 2010) the approach was used in the context of product reviews and in (Kennedy and Inkpen, 2006) in the context of film reviews. Classifier based opinion mining techniques have been shown to perform well, however their usage features the disadvantage that a pre-labelled training set is required. The resource needed to build such a training set is often prohibitive. A solution is the lexicon based approach where sentiment lexicons are used to estimate the sen-

timent value/score and polarity (attitude) expressed within documents in a corpus by first identifying subjective words (words that convey feelings or judgement) and then “looking up” the identified words in a sentiment lexicon to obtain sentiment values and polarities (positive or negative) for each word. These values and polarities can then be used to predict the overall polarity (attitude) for each document in the corpus (Esuli and Sebastiani, 2006; Denecke, 2009; Montejo-Raez et al, 2012; Ohana and Tierney, 2009; Salah et al, 2013a).

The most commonly used sentiment lexicon is SentiWordNet 3.0 which has the key advantage, over other such lexicons, that it covers a larger number of words (117659 words). The problem with such general purpose lexicons is that they tend to not operate well with respect to specific domain corpora, because of the use of special purpose words (reserved words) and/or domain specific style and language that may be a feature of specialised domains such as the political debate domain. For example, given a specific domain, certain words and phrases may be used in a different context than their more generally accepted usage, in which case the words and phrases may reflect different sentiments than those that would be normally expected. A solution is to use domain specific lexicons, however these tend not to be readily available and thus have to be generated. There are two approaches to generating such domain specific lexicons: (i) direct generation and (ii) adaptive generation (Salah et al, 2013b). The first, as the name suggests, is founded on the idea of generating the desired domain-specific lexicon directly using the biased occurrence of words in a given pre-labelled training corpus (thus obviating the claimed advantage of lexicon-based opinion mining approaches over classification based approaches that a training set is not required). The second approach is founded on the idea of using an existing general purpose lexicon, such as SentiWordNet 3.0, and adapting this so that it becomes a domain specific lexicon, again using pre-labelled training data.

In the context of political sentiment mining we thus have three potential approaches that we can adopt: (i) classification based, (ii) generic lexicon based and (iii) domain specific lexicon based (two techniques, direct and adaptive). One of the objectives of this paper is to compare the operation of these three approaches in the context of predicting the attitude (for or against a motion) of individual speakers in a political debate. The comparison is conducted by applying the approaches to the concatenated speeches of individual Members of Parliament (MPs) conducting political debates within the UK House of Commons so as to determine the attitude of the speakers. More generally we are also interested in whether the analysis of such speeches is a good indicator of attitude or whether some alternative indicator, such a political affiliation, is a better indicator of attitude.

The rest of this paper is structured as follows. Section 2 provides some background on sentiment lexicons and machine learning sentiment classification, and reviews some relevant previous work. Section 3 presents the UK House of Commons dataset used for evaluation purposes within this paper. Sections 4, 5 and 6 describe the three potential approaches that we can adopt and include some innovations proposed by the authors. The comparison of the approaches is then presented in Section 7 together with consideration of whether political debate speeches are a good indicator of speaker attitude, or whether some other predictor (such as party affiliation) is a better predictor. The main findings of the work and some concluding observations are then presented in Section 8.

## 2 Previous work

In this previous work section we provide some background concerning the three approaches to sentiment mining that are of interest with respect to the work described in this paper. We commence in Sub-section 2.1 with the classification based approach and then go on, Sub-section 2.2, to consider the lexicon based approaches (generic and domain specific). The section is completed with a review of recent work on sentiment analysis in the political domain.

### 2.1 The classifier based approach to sentiment mining

In the classification (machine learning) based approach to sentiment mining a pre-labelled training corpora (exhibiting prior knowledge) is used to learn a “classifier” using some established supervised learning mechanism. The training data comprises a collection of ordered pairs  $\langle a, c \rangle$  where  $a$  is an instance (observation) comprised of a set of attribute values and  $c$  is a known class label for that instance taken from the set  $C$ . Once the classifier has been generated it can be used to assign documents to the “fittest” class; essentially performing a mapping  $a_i \rightarrow c_i$  where  $c_i \in C$  (the set of known class labels). It has been argued that classification based approaches in political opinion mining tend to outperform lexicon-based approaches (Grijzenhout et al, 2010). However, the need for appropriate training data is a limiting factor, and the learning process is highly dependent on the quality of the prior knowledge (historical data) available.

## 2.2 The lexicon based approach to sentiment mining

Sentiment lexicons are lexical resources used to support sentiment mining. More specifically they are used to assign a sentiment value (or score) and a polarity (or orientation) to a word. A sentiment value is a numeric value indicating some degree of subjectivity. The polarity (positive or negative) of a word is an indicator of whether the word expresses assent or dissent with respect to some object or concept. Consequently, document polarity can be judged by counting the number of positive and negative subjective words, summing their sentiment values and then calculating the difference. The result represents the attitude (positive or negative) of the document. Relatively small sized sentiment lexicons, which are built manually, can be extended by applying lexical induction techniques that exploit the semantic relationships between terms and their synonyms and antonyms, or by measuring term similarities in large corpora. As mentioned in Section 1 two types of sentiment lexicons can be used in the context of sentiment mining: (i) generic (domain-independent) and (ii) dedicated (domain-specific) sentiment lexicons. More detail concerning these two types of lexicons will be discussed below in Sub-sections 2.2.1 and 2.2.2 respectively.

### 2.2.1 Generic lexicon based sentiment mining

The most commonly used generic (topic-independent) sentiment lexicon is the “off-the-shelf” SentiWordNet 3.0<sup>1</sup> sentiment lexicon (Baccianella et al, 2010), which is founded on WordNet 3.0<sup>2</sup>. WordNet is a large lexical repository of English words grouped into sets of cognitive synonyms called *synsets* expressing distinct concepts. Synsets are interlinked by means of conceptual-semantic and lexical relations. SentiWordNet 3.0 is an extension of SentiWordNet 2.0 (based on WordNet 2.0) which in turn is derived from SentiWordNet 1.0 (based on WordNet 1.0) (Esuli and Sebastiani, 2006). SentiWordNet 3.0 associates to each synset  $s$  of WordNet a set of three scores:  $Pos(s)$  (“positivity”),  $Neg(s)$  (“negativity”),  $Obj(s)$  (“neutrality” or “objectivity”). The range of each score is  $[0, 1]$  and for each synset  $s$ ,  $Pos(s) + Neg(s) + Obj(s) = 1$ . Table 1 presents some statistics with respect to a number of popular sentiment lexicons, including SentiWordNet 3.0 (Ohana and Tierney, 2009). From the table it can be seen that, out of the four lexicons listed, SentiWordNet 3.0 has the key advantage of covering the largest number of words.

---

<sup>1</sup> SentiWordNet 3.0 is accessible at [sentiwordnet.isti.cnr.it](http://sentiwordnet.isti.cnr.it).

<sup>2</sup> WordNet is accessible at <http://wordnet.princeton.edu/>.

**Table 1** Coverage of SentiWordNet 3.0 compared to other (manually built) sentiment lexicons (Ohana and Tierney, 2009).

Sentiment lexicon	Generation	Total num. sentiment bearing terms
<b>SentiWordNet 3.0</b>	<b>Automatically</b>	<b>117659</b>
Subjectivity Clause Lexicon	Manually	7650
General Inquirer	Manually	4216
Grefenstette	Manually	2258

### 2.2.2 Domain specific lexicon based sentiment mining

As noted above, sentiment analysis using generic sentiment lexicons is a challenging process in the context of topic-dependent domains (Thelwall and Buckley, 2013). In such cases it is desirable to use dedicated domain specific sentiment lexicons. However, the main issue with the usage of such dedicated lexicons is that they are frequently not readily available and thus have to be specially generated, a process that may be both resource intensive and error prone.

As noted in the introduction to this paper two approaches may be identified to generating specialised (dedicated) lexicons for domain specific sentiment analysis: (i) creating a new dedicated lexicon or (ii) adapting an existing generic lexicon. Both techniques use labelled corpora (training data) from a specific domain. An example of the first technique (creating a new dedicated lexicon) can be found in (Birla et al, 2011) who proposed a semi-automated mechanism to extract domain-specific health and tourism words from noisy text so as to create a domain-specific lexicon. Examples of the second technique (adapting an existing general lexicon) can be found in (Demiroz et al, 2012) and (Choi and Cardie, 2009). In (Demiroz et al, 2012) a simple algorithm was proposed to adapt a generic sentiment lexicon to a specific domain by investigating how the words from the generic lexicon are used in the specific domain context in order to assign new polarities to these words. In (Choi and Cardie, 2009) Integer Linear Programming was used to adapt a generic sentiment lexicon into a domain-specific lexicon; the method combined the relations among words and opinion expressions so as to identify the most probable polarity of lexical items (positive, negative, or neutral or negator) for the given domain. Interested readers are referred to (Salah et al, 2013b) for further discussion regarding the generation of domain specific lexicons for sentiment mining.

There is also reported work that combines the two techniques (adapting the sentiment scores of the terms in the base lexicon and additionally appending new domain words to extend the base lexicon). For example (Weichselbraun et al, 2011) created a domain-specific sentiment lexicon using crowd-sourcing for assigning sentiment scores to sentiment terms and then automatically extending an initially pro-

posed base lexicon using a bootstrapping process to add new sentiment indicators and terms. The lexicon is then customised according to some specific domain. The evaluation conducted indicated that the created lexicon outperforms a generic sentiment lexicon (the General Inquirer Sentiment Lexicon<sup>3</sup>). Further reported work concerned with the “dual approach” to generating domain-specific lexicons can be found in (Qiu et al, 2009; Lau et al, 2011; Ringsquandl and Petković, 2012).

Sentiment score for each term in a lexicon can be calculated either by: (i) investigating the biased occurrence of the term with respect to the class labelled documents (“positive” and “negative”), (ii) utilising the semantic, contextual or statistical relationships between terms (words) in an input domain corpus, or (iii) learning a classifier to assign sentiment polarity to terms. In the context of the calculation of sentiment scores with respect to specific domains, (Zhang and Peng, 2012) proposed a method to calculate the sentiment score of each word or phrase in different domains and use these scores to quantify sentiment intensity. (Thelwall and Buckley, 2013) proposed two approaches to improve the performance of polarity detection using lexical sentiment analysis with respect to social web applications, focusing on specific topics (such as sport or music). The two approaches were: (i) allowing the topic mood to determine the default polarity for false-neutral expressive text, and (ii) extending an existing generic sentiment lexicon by appending topic-specific words. The mood method slightly outperformed the lexical extension method. On the other hand it was found to be very sensitive to the “mood base” used, thus it was necessary to analyse the corpus first in order to choose an appropriate mood base relative to the corpus. Both methods require human intervention to either annotate a corpus (mood method) or to select terms (lexical extension).

### 2.3 Related work on sentiment analysis in the political domain

In (Grijzenhout et al, 2010) two sentiment mining techniques were considered, based on two different models to automatically identify the subjectivity and orientation of text segments, to retrieve political attitudes or viewpoints from Dutch parliamentary publications. The outcomes were then compared with a manually compiled and annotated “gold standard”. The first of the two techniques used machine learning classifiers (Naive Bayes, Support Vector Machine SMO, BK1 nearest neighbour and ZeroR), while the second was a dictionary (lexicon) based technique that used a subjectivity lexicon. Despite the fact

---

<sup>3</sup> The General Inquirer Sentiment Lexicon was built using the sentiment information contained in the General Inquirer (see (Stone et al, 1966) for more information about General Inquirer).

that the machine learning approach outperformed the lexicon-based approach the results indicated that both opinion mining techniques were applicable for investigating subjectivity and sentiment polarity in Dutch political semi-structured transcripts. (Rissland, 1999) manually surveyed and discussed different types of arguments made in the short (nearly one-minute) speeches given during the last hour of the debate (hearing) held within the United States House of Representatives in December 1998 on the articles of impeachment of President Clinton. The author demonstrated that these short speeches featured a reduced structure compared to the longer speeches more usually made in the House of Representatives which tended to be more structured and coherent. In (Thomas et al, 2006) work was described on determining, using the transcripts of U.S. Congressional floor debates, the degree of agreement between opinions expressed by speakers' speeches supporting or opposing proposed legislation. By utilising information about the inter-document relationships between speeches (in particular, whether two speeches belonged to the same speaker, or whether they shared similar "content") it was demonstrated that this improved a "support" versus "oppose" classification over the classification of speeches in isolation. The "support"/"oppose" classification and its usefulness in debate visualisation is also argued for in (Birnbaum, 1982) who identified a number of frequent patterns of interaction in argument.

### 3 The UK House of Commons political debates corpus

To act as a focus for the work described in this paper UK House of Commons debates were used. Both houses in the UK parliament, the House of Commons and the House of Lords, reach their decisions by debating and then voting with either an "Aye" or a "Nay" at the end of each debate. Proceedings of the Commons Chamber are published on-line in XML format three hours after they take place (at [TheyWorkForYou.com](http://TheyWorkForYou.com)). Figure 1 shows an extract from a debate transcript concerning the "Enterprise and Regulatory Reform Bill, Tuesday 26 June 2012"<sup>4</sup> debate. The highlighted text indicates MPs who voted "Aye" at the end of the debate while the unhighlighted text indicates MPs who voted "Nay". The advantage offered by this collection is that the outcome of the debates are known and thus this collection can be used evaluate the veracity of the outcomes of the application of opinion mining techniques such as those considered in this paper.

---

<sup>4</sup> Hansard source citation: from Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c225 to Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c226.



**David Mowat** (Warrington South, Conservative)  
 I do. My point was that, if we were going through the catalogue of achievement, in 2010 we were 25th out of 27, notwithstanding the points that have been made. All I am saying is that the Green investment bank as designed is part of a project to catch up on that position. The hon. Gentleman talks about slipping down the league table; we have only Cyprus and Malta to go. Let us be clear about where we are starting from.

Hansard source (Citation: Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c225)



**Iain Wright** (Hartlepool, Labour)  
 The point I am trying to make is that we do have a competitive advantage in a number of sectors in the green economy. We need to take advantage of that. Other nations realise that the green economy is a driver for economic growth. There needs to be a greater sense of urgency. I want to work closely with the Minister on this. There is a window of opportunity that is closing faster than the Committee is considering the clauses of the Bill. We need to act fast and boldly, because our successors will think about the subject, debate it in the House in 2020 or 2030, and say, as we did in the 1980s in relation to onshore wind technology, "We were market leaders in this, but lack of Government support meant that we slipped behind other nations, and the likes of Germany and Denmark are taking our place." That should not be allowed to happen. We should ensure that we are at the forefront of the global green industrial revolution.

Hansard source (Citation: Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c225)



**Chris Ruane** (Vale of Clwyd, Labour)  
 Why and how is it that the Government claim to be the greenest Government ever? Does my hon. Friend agree that if we believed that, we would be the greenest Opposition ever?

Hansard source (Citation: Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c226)



**Iain Wright** (Hartlepool, Labour)  
 That is a good point. I do not think that we are as naive as that. The Minister will correct me if I am wrong, and I do not want to mislead the Committee, but it is interesting that, in the two years or so that the Prime Minister has been in office, he has not made a single speech about the environment. I think that that is correct. I know that the Committee is anxious to hear about the Lib Dem manifesto. Is the Minister keen to talk about it? If so, I am more than happy to give way.

Hansard source (Citation: Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c226)



**Neil Carmichael** (Stroud, Conservative)  
 I am not the Minister, and that is not a document that I know as well as others do, but does the shadow Minister agree that the fact that we are introducing the Green investment bank is a signal of our commitment to being green? What about the green deal, which is another huge step in the right direction of greening the environment? Various other measures that we have taken include the Energy Act 2011. Does he agree that they are joined-up, consistent and emblematic of a green Government?

Hansard source (Citation: Enterprise and Regulatory Reform Bill Deb, 26 June 2012, c226)

**Fig. 1** Fragment of a UK House of Commons debate as published on the They-WorkForYou.com www site.

QDAMiner<sup>4</sup><sup>5</sup> was used to extract the desired textual information from the XML debate records. In this manner the authors obtained the speeches associated with 29 different UK House of Commons debates held between August 2012 to March 2013 and extracted the desired

<sup>5</sup> <http://provalisresearch.com>

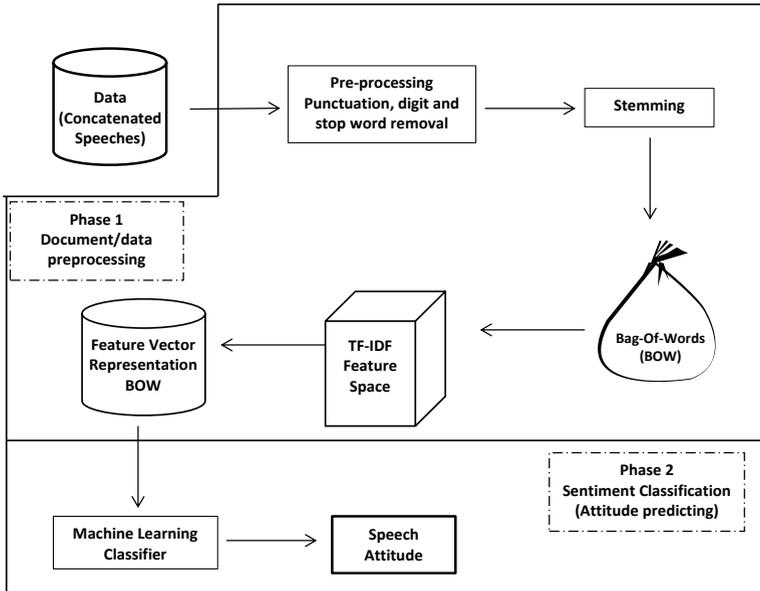
**Table 2** Debate Dataset statistical overview.

Debate ID	Aye Sp.	Nay Sp.	Total Sp.	Min length	Max length	Total words	Avg length
D1	51	55	106	50	4881	72376	682.792
D2	38	50	88	51	4974	82506	937.568
D3	29	22	51	51	4822	40766	799.333
D4	41	40	81	50	4843	65523	808.926
D5	37	36	73	51	4989	65214	893.342
D6	54	53	107	55	4804	50146	468.654
D7	39	43	82	50	4880	60474	737.488
D8	21	2	23	62	2878	22510	978.696
D9	39	40	79	50	4917	81095	1026.519
D10	40	6	46	51	4680	47207	1026.239
D11	35	48	83	54	4928	74375	896.084
D12	32	19	51	54	4951	75525	1480.882
D13	6	25	31	63	4846	39662	1279.419
D14	34	31	65	55	4896	77766	1196.400
D15	18	3	21	66	4872	34900	1661.905
D16	66	28	94	53	4804	59589	633.926
D17	55	51	106	51	4849	74703	704.745
D18	47	47	94	51	4915	72770	774.149
D19	42	40	82	50	4926	98845	1205.427
D20	45	41	86	50	5041	70457	819.267
D21	28	12	40	50	4999	37866	946.650
D22	80	40	120	51	4890	88518	737.650
D23	44	34	78	51	4817	74880	960.000
D24	44	60	104	54	4986	113701	1093.279
D25	33	19	52	50	4710	60114	1156.038
D26	2	8	10	86	4929	19107	1910.700
D27	37	29	66	51	4967	63392	960.485
D28	54	47	101	50	4961	81490	806.832
D29	28	20	48	58	4863	32519	677.479
<b>MIN</b>	2	2	10	50	2878	19107	468.654
<b>MAX</b>	80	60	120	86	5041	113701	1910.700
<b>AVG</b>	38.586	32.724	71.310	54.103	4821.310	63379.172	974.513
<b>Total</b>	1119	949	2068	1569	139818	1837996	28260.876

textual information. For each debate the speeches associated with the same MP were concatenated together. Concatenated speeches by MPs who did not vote were ignored; as were speeches that contained fifty words or less, as it was conjectured that no valuable sentiment attitude could be associated with such short speeches. The dataset comprised 2,086 concatenated speeches (1,119 speeches made by speakers who voted Aye and 949 speeches made by speakers who voted Nay) associated with 553 distinct Members of Parliament (MPs) belonging to 10 distinct political parties. The speeches comprised a total of 1,837,996 words or uni-grams (17,893 unique words after stemming and stop-word removal or 31,259 unique words after only stop-word removal). Note that the number of concatenated speeches featured in a debate also equated to the number of MPs taking part. Some statistics concerning this dataset are presented on Table 2. From the table, the average number of words in a concatenated speech is 975 and in a whole debate is 63,379. The average number of Aye and Nay speeches (39 and 33 respectively) is reasonably balanced.

## 4 Political sentiment mining using classification

In this and the following three sections we present our implementations of the three opinion mining approaches of interest in the context of mining the UK House of Commons political debates, starting with the classification based approach. An overview of the proposed speaker attitude classification process is presented in Figure 2. The input is the set of concatenated speeches that make up a single debate, the output is a set of attitude labels one per concatenated speech. More formally the input is a set of  $n$  concatenated speeches  $S = \{s_1, s_2, \dots, s_n\}$ , and the output is a set of attitude class labels  $C = \{c_1, c_2, \dots, c_n\}$  taken from the set  $\{positive, negative\}$  such that there is a one-to-one correspondence between the elements in  $S$  and  $C$ . The process encompasses two stages: (i) preprocessing and (ii) attitude prediction. Each of these stages is described in more detail in the following two Sub-sections.



**Fig. 2** The classification based approach to sentiment mining.

## 4.1 Preprocessing

In terms of text processing each concatenated speech, associated with a speaker (MP), can be conceptualised as a document. During the preprocessing stage all upper-case alphabetic characters are first converted to lower-case letters followed by numeric digit removal. This is followed by a tokenisation process, the “breaking up” of the document set contents into sets of primitive components called *tokens*, words in our case, which are identifiable by white space separators and/or punctuations (. , ; : ’ ” ( ) ? !). The resulting tokens are then indexed to form an initial Bag-Of-Words ( $BOW = \{t_1, t_2, \dots, t_{|BOW|}\}$ ). The next step is to reduce the size of the BOW by removing “stop words”. Stop words are words which are not expected to convey any significant meaning in the context of sentiment analysis, for example words such as “the” and “and” (Chim and Deng, 2008; Hariharan and Srinivasan, 2008; Poomagal and Hamsapriya, 2011). Each document will now be represented by some subset of the BOW. Given a specific domain there will also be additional words, other than stop words, that occur frequently. In the case of our parliamentary debates words like: “hon.”, “house”, “minister”, “government”, “gentleman”, “friend” and “member” are all very frequent words. For similar reasons as for stop word removal these domain specific words are also removed. This was done by appending them to the stop-words list. The names of all the members of parliament, political parties and constituencies were also added to the stop-word list.

The size of our BOW is then further reduced by applying stemming. Stemming is concerned with the process of deriving the “stem” of a given word by removing the added affixes so that “inflated” words that belong to the same stem (root) will be “counted together” (Hariharan and Srinivasan, 2008). For example “compute”, “computes”, “computer”, “computed”, “computation” and “computing” will all be reduced to the common stem “compute”.

On completion of the preprocessing and stemming stages the resulting BOW defines a feature spaces from which sets of feature vectors can be generated. The feature vector elements hold term weightings. The most widely used mechanism for generating term weightings, and that adopted with respect to the work described in this paper, is the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme which aims to “balance out the effect of very rare and very frequent” terms in a vocabulary (Kuhn et al, 2007). TF-IDF also tends to reflect the significance of each term by combining local and global term frequency (Li et al, 2009). TF-IDF can be defined as follows:

$$w_{ij} = TFIDF(i, j) = tf(i, j) \cdot \left( \log \frac{N}{df(j)} \right) \quad (1)$$

where: (i)  $tf(i, j)$  is the frequency of term  $j$  in document  $d_i$  (local weight for the term), (ii)  $N$  is the total number of documents in the corpus (concatenated speeches in the debate), and (iii)  $df(j)$  is the number of documents (speeches) containing term  $j$  (global weight for the term).<sup>6</sup> Table 3 shows the document frequency counts for a number of example terms taken from our parliamentary debate collection. The table also shows the document count with respect to documents (speeches) where the MP in question voted “Aye” and where the MP voted “Nay”. The final column gives the document frequency difference between the number of “Aye” and “Nay” counts. Inspection of this final column clearly indicates that some terms can be associated with an “Aye” vote, while other terms can be associated with a “Nay” vote. For example *cuts* is associated with an Aye vote while *european* is associated with a Nay vote (during the period when our political speeches were collected, August 2012 to March 2013, a right of centre political party was in government in the UK who had a tendency to favour tax cuts and oppose European integration).

**Table 3** Document Frequency (DF) values (indicating biased occurrences) associated with selected terms occurring in a set of debates with respect to MPs who voted “Aye” and “Nay”.

Term	DF (Aye)	DF (Nay)	DF (Total)	Difference	Term	DF (Aye)	DF (Nay)	DF (Total)	Difference
people	406	338	744	68	timetable	23	23	46	0
cuts	87	38	125	49	taxpayer	11	29	40	-18
change	154	111	265	43	generous	10	28	38	-18
worse	52	17	69	35	fully	34	53	87	-19
simply	101	70	171	31	sustainable	11	33	44	-22
care	69	39	108	30	funding	41	64	105	-23
confidence	60	31	91	29	improve	40	63	103	-23
recession	42	13	55	29	assure	34	59	93	-25
women	64	36	100	28	inherited	9	38	47	-29
military	42	16	58	26	previous	101	131	232	-30
hope	136	120	256	16	raises	8	38	46	-30
existence	15	0	15	15	reduce	38	73	111	-35
wonderful	24	10	34	14	encourage	30	72	102	-42
deep	21	7	28	14	european	59	105	164	-46

Thus at the end of the pre-processing phase the input collection of speeches will be represented using the vector space model such that each speech is described by a feature vector. More formally a speech  $i$  is represented as a vector  $V_i = \{w_{i1}, w_{i2}, \dots, w_{iz}\}$  where  $w_{ij}$  is the TF-IDF value for term  $j$  in speech  $i$ . It should also be noted that each element in  $V_i$  corresponds to a term in the BOW. Again, we will indicate the list of terms associated with feature vector  $V_i$  using the notation  $T_i = \{t_{i1}, t_{i2}, \dots, t_{iz}\}$ . Thus we have a set of feature vectors  $V = \{V_1, V_2, \dots, V_z\}$  and a set of term lists  $T = \{T_1, T_2, \dots, T_z\}$  with a one-to-one correspondence between the two.

<sup>6</sup> Alternative schemes to TF-IDF include: Term Frequency (TF), Document Frequency (DF), Term Strength (TS) and Term Contribution (TC).

## 4.2 Attitude prediction using machine learning classifiers

Once the input data has been translated into the feature vector format, whereby the concatenated speeches for each speaker are defined by a subset of words contained in the BOW, classification can be applied to determine each speakers “attitude” (positive or negative). The idea is founded on the natural assumption that the nature of the speeches made by MPs can be used as clues as to how they are going to vote (although this assumes that speakers do not “change their mind” during a debate). To this end we require a classifier. Classifier generation is a supervised machine learning mechanism (as noted previously in Section 2.1) that requires pre-labelled training data (something which we would only have with respect to historical data). In our case we used the known vote associated with each speaker in the dataset as the label. Any number of different classifier generation techniques could have been adopted, however in the context of the comparison presented later in the paper, a number of classifier generator techniques available within the Weka-3.6 workbench<sup>7</sup> were considered: (i) Naive Bayes, (ii) Support Vector Machine SMO, (iii) J48 decision tree learner, (iv) JRip rules-based classier, (v) IBk nearest neighbour classier and (vi) ZeroR. Once a classifier has been generated it can be evaluated by applying in to pre-labelled test data and the labels produced compared with the known labels. Provided that the generated classifier is found to be sufficiently effective. Note that the training data used to generate the classifier, the data used to evaluate it and the data to which it is to be applied, all have to be preprocessed in the same manner.

## 5 Political sentiment mining using generic sentiment lexicons

Given a new text which we wish to classify as expressing either a “positive” or a “negative” opinion, the subjective words in the text act as sentiment indicators. However, subjective word identification is a challenging process because of the complexity of natural language. One solution is to use sentiment lexicons, which can be used to look-up words to firstly identify subjective words (as opposed to objective words) and secondly to determine the degree of sentiment and polarity (positive or negative) associated with the identified subjective words. This information can then be used to make a judgement about the overall sentiment represented by a text. The main idea is to combine the subjective word-level sentiment values to give a whole document sentiment value. As already noted above there are two approaches to

---

<sup>7</sup> <http://www.cs.waikato.ac.nz/ml/weka/downloading.html/>

sentiment mining using sentiment lexicons, we can either use an off-the-shelf generic lexicon or use a domain specific lexicon. This Section is directed at the generic approach, the following section considers the domain specific approach. In both cases, part-of-speech tagging is performed first to assign a part-of-speech tag for each word in the input text as described in Sub-section 5.1. The second step is text preprocessing. This preprocessing is similar to that proposed with respect to the classification based technique and is described in Sub-section 5.2 below. Once the data has been pre-processed the attitude prediction (mining) phase can be commenced, this is described in Sub-section 5.3

### 5.1 Part Of Speech Tagging (POST)

Part-Of-Speech Tagging (POST) is a process whereby each word in a given text is assigned a POS tag according to its context in the sentence or a phrase in which it is used (Bellegarda, 2010). With respect to sentiment mining POST is important because many related words (for example “suffice”, “sufficiency”, “sufficient” and “sufficiently”), which have different POS tags, will typically have different sentiment scores. POST is also significant with respect to sentiment mining because it allows for Word Sense Disambiguation (WSD), the process of dealing with the polysemy problem (different meanings for the same word) by discriminating the proper “semantic” sense of a word in a specific context or circumstance (Wilks and Stevenson, 1998). At the end of the POST step we will have a list of terms  $T = \{t_1, t_2, \dots, t_m\}$  each associated with a POS tag  $post_i$ , thus a set of pairs  $\langle t_i, post_i \rangle$ .

### 5.2 Preprocessing

Once POS tagging is complete the pre-processing stage can be commenced. As in the case of the classification based approach the first steps in the preprocessing are tokenisation and stop word removal. Stemming was not used in the context of the lexicon based approaches because words like “suffice”, “sufficiency”, “sufficient” and “sufficiently” will have different Part Of Speech (POS) tags and will consequently have different sentiment scores. When stemming is applied, these words will be reduced to a single word (stem) and thus share the same sentiment score therefore possibly losing the more appropriate individual sentiment values. Instead a lemmatisation approach was adopted. Lemmatisation is different from stemming in that the aim is to reduce a given word to its “conventional standard form” instead of its root or stem form. For example all verbs would be converted to their infinitive form and all nouns to their singular form (Amine et al, 2010).

On completion of tokenisation, stop-word removal and lemmatisation a BOW representation was again used, as in the case of the classification based approach, however in this case each word in the BOW will be linked to a POS tag. Thus each document (speech) will now be represented by some subset of the *BOW* which in turn is translated into a feature vector form.

### 5.3 Attitude detection using sentiment lexicons

Given our feature vector representation sentiment analysis is applied to the terms in each vector to determine the attitude reflected by the vector and consequently the document (concatenated speech) it represents. The “sentiment” score (value) associated with each term in  $t_i$  in feature vector  $S_i$  is obtained by “looking up” the term in a sentiment lexicon. As noted previously in Section 2 a sentiment score is a numeric value indicating some degree of subjectivity. The orientation of a word is an indicator of whether a word expresses assent or dissent with respect to some object or concept. Consequently document polarity can be judged by counting the number of positive and negative terms and calculating the difference. The resulting polarity then describes the attitude reflected by the document.

With respect to the generic lexicon based approach SentiWordNet was used. SentiWordNet assigns a positive and a negative score (ranging from 0.0 to 1.0) to each synset that exists in WordNet so as to generate polarity scores. The synsets in SentiWordNet 3.0 were broken down into single terms in order to produce a list of terms which are then used to retrieve the corresponding score. Terms which originate from the same synset are taken to have the same sentiment score. However, if a term features in different synsets then: (i) if the different forms of the term in the different synsets have different grammatical tagging (POS tag), then the “word-sense distinction” is resolved simply by considering the different POS tags of the term (as suggested in (Wilks and Stevenson, 1998)) and thus it is split into distinguished terms; (ii) if the term has the same grammatical tagging in the different synsets, the highest sentiment score is selected.

More formally, from the above, the accumulative sentiment score  $st_i$  associated with a speech  $i$  is computed using:

$$st_i = \sum_{j=1}^{j=z} (\text{Lex}(term_j) \times w_{ij}) \quad (2)$$

where: (i)  $term_j$  is a term in the the feature vector representing speech  $i$ ; (ii)  $\text{Lex}$  is a function that returns the sentiment score ( $-1.0 \leq \text{Lex}(term_j) \leq +1.0$ ) for each  $term_j$ , from the adopted sentiment lexicon, where the score is the summation of the term’s positivity and

negative scores (positive and negative values); (iii)  $z$  is the number of terms in the given feature vector and (iv)  $w_{ij}$  is the occurrence count for term  $j$  in feature vector  $i$ . The occurrence count can be a true frequency count of the number of times  $term_j$  appears in document  $i$ , or simply a binary value (1 or 0) to indicating the present or absence of the term. In the upcoming sections we refer to these two techniques using the labels *TF* and *Binary* respectively. The attitude (class label) for each document (speaker)  $i$  is then determined according to  $st_i$ . To this end the class label set is  $\{positive, negative, objective, neutral\}$  where: (i) *positive* indicates a positive text (for the motion in the case of our political debates), (ii) *negative* indicates a negative text (against the motion), (iii) *objective* indicates that no sentiment scores were found and (iv) *neutral* that the sentiment scores negate each other. In practice it was found that the last two class labels are rarely encountered. Algorithm 5.1 describes the attitude identification process. The algorithm loops through the input set of speeches, represented in terms of the sets  $S$  and  $T$  (see end of Section 5.2), a sentiment score for each speech is calculated from lines 7 to 23, the attitude from lines 24 to 38.

## 6 Political sentiment mining using domain specific sentiment lexicons

Political sentiment mining using domain specific lexicons operates in a similar manner to that using generic lexicons with the exception that dedicated lexicons are used. The challenge is obtaining the required specialist lexicons. As discussed in Sub-section 2.2.2 two approaches to generating domain-specific sentiment lexicons can be identified: (i) direct generation and (ii) adaptive generation. In this context the authors have proposed techniques for both direct and adaptive domain specific lexicon generation. For completeness these two techniques are included in this section. In both cases the input is a set of  $n$  binary labelled parliamentary speeches (documents)  $D = \{d_1, d_2, \dots, d_n\}$ . The labels are drawn from the set  $\{positive, negative\}$ . The output in both cases is a lexicon where each term is encoded in the form of a set of tuples  $\langle t_i, post_i, s_i \rangle$ , where  $t_i$  is a term that appears in the document collection  $D$ ,  $post_i$  is the part-of-speech tag associated with term  $t_i$  and  $s_i$  is the associated sentiment score. Both domain-specific lexicon generation approaches comprise four steps: (i) part-of-speech tagging (to identify the POS tags), (ii) document preprocessing, (iii) sentiment score ( $s_i$ ) and polarity ( $post_i$ ) calculation and (iv) lexicon generation (see Figure 3). Each of these steps is described in more detail in the following four Sub-sections. With respect to the evaluation described later in this paper, these two techniques were used to create

**Algorithm 5.1** Attitude identification using sentiment lexicon

---

```

1: INPUT: Sentiment Lexicon  $Lex$ , set of sets of terms  $T \subset BOW$ , set of feature
  vectors  $S$ 
2: OUTPUT: Set of Attitudes labels  $A = \{a_1, a_2, \dots, a_z\}$ 
3:  $PosCount = 0$ 
4:  $NegCount = 0$ 
5:  $PosScore = 0$ 
6:  $NegScore = 0$ 
7: for all  $T_i \in T^1$  do
8:   for all  $t_{ij} \in T_i$  do
9:     if  $t_{ij} \in Lex$  then
10:       $score_{ij} = Lex(t_{ij}) \times w_{ij}$ 
11:     else
12:       $score_{w_{ij}} = 0$ 
13:     end if
14:     if  $Score_{ij} > 0$  then
15:        $PosCount = PosCount + w_{ij}$ 
16:        $PosScore = PosScore + Score_{ij}$ 
17:     else if  $Score_{w_{iu}} < 0$  then
18:        $NegCount = NegCount + w_{ij}$ 
19:        $NegScore = NegScore + Score_{ij}$ 
20:     else[ $Score_{ij} = 0$ ]
21:       DO NOTHING
22:     end if
23:   end for
24:   if  $PosCount = 0 \wedge NegCount = 0$  then
25:      $a_i = Objective$ 
26:   else if  $PosScore > NegScore$  then
27:      $a_i = Positive$ 
28:   else if  $NegScore > PosScore$  then
29:      $a_i = Negative$ 
30:   else[ $PosScore = NegScore$ ]
31:     if  $PosCount > NegCount$  then
32:        $a_i = Positive$ 
33:     else if  $NegCount > PosCount$  then
34:        $a_i = Negative$ 
35:     else[ $PosCount = NegCount$ ]
36:        $a_i = Neutral$ 
37:     end if
38:   end if
39: end for

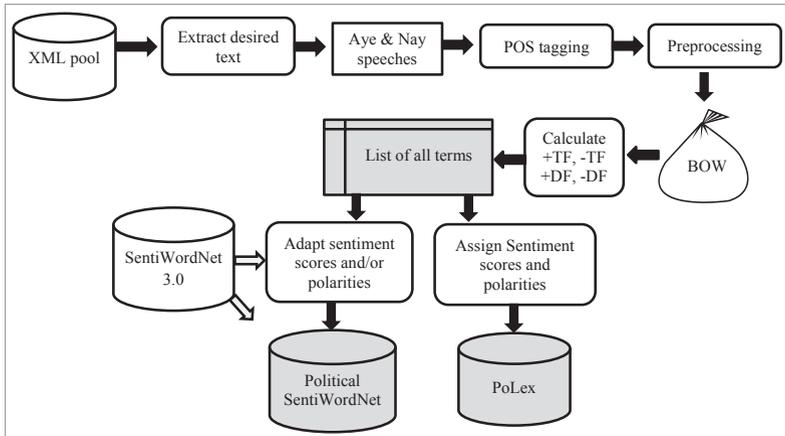
```

---

two political-domain sentiment lexicons from our UK House of Commons political debate data: PoLex produced using direct generation and PoliSentiWordNet produced using adaptive generation.

### 6.1 Part Of Speech Tagging (POST)

The first step with respect to the two lexicon generation approaches is POS tagging so that each word in the input is assigned a particular POS tag to produce a list of terms  $T = \{t_1, t_2, \dots, t_m\}$  each associated



**Fig. 3** Domain-specific lexicon direct creation and adaptation.

with a POS tag  $post_i$ . The practice of POST was discussed in detail in Sub-section 5.1 above.

## 6.2 Preprocessing

The pre-processing step commences with the conversion of all uppercase alphabetic characters to lower-case, this is then followed by punctuation mark and numeric digit removal. Next, given our list of terms  $T$  from the previous step, we create a Bag-Of-Words (BOW) representations for all  $t_i$  in  $T$  (all the terms in the input document collection  $D$ ). Each term  $t_i$  in the BOW is defined using a 6-tuple of the form  $\langle t_i, post_i, tf_i^+, tf_i^-, df_i^+, df_i^- \rangle$ , where (i)  $t_i$  is the term of interest (term number  $i$ ); (ii)  $post_i$  is the associated POS tag as identified in the previous step, (iii)  $tf_i^+$  is the associated term frequency (number of occasions that the term  $t_i$  appears in a text collection) with respect to texts that display a positive attitude (“Aye” labelled texts in the case of our political speeches), (iv)  $tf_i^-$  is the associated term frequency with respect to texts that display a negative attitude (“Nay” labelled texts in the case of our political speeches), (v)  $df_i^+$  is the associated document frequency (number of texts in which  $t_i$  appears) with respect to texts that display a positive attitude (“Aye” labelled documents) and (vi)  $df_i^-$  is the associated document frequency with respect to texts that display a negative attitude (“Nay” labelled documents).

### 6.3 Determining the sentiment score and polarity of each term

On completion of the pre-processing step, sentiment scores (sentiment weightings) are calculated with respect to each term contained in the generated BOW so far. The TF-IDF weighting value  $W_{ij}$  for a term  $t_i$  in a text  $j$  is obtained using:

$$W_{ij} = TF - IDF = tf_{i,j} \cdot \left( \log_2 \frac{n}{df_i} \right) \quad (3)$$

where: (i)  $tf_{i,j}$  is the frequency of term  $t_i$  in document (speeches)  $j$  (thus the local weight for the term), (ii)  $n$  is the total number of documents in the corpus (concatenated speeches in the debate), and (iii)  $df_i$  is the number of documents (speeches) containing term  $t_i$  (thus the global weight for the term). A disadvantage of TF-IDF, in the context of opinion mining, is that it does not reflect a term's sentiment tendency (orientation) and thus for our purposes we need either an alternative sentiment intensity weighting scheme or an alternative form of the TF-IDF scheme that takes into consideration the situation where a term  $t_i$  appears in both positive and negative documents. With respect to the latter the  $\Delta$ TF-IDF provides "an intuitive general purpose technique to efficiently weight word scores" (Martineau and Finin, 2009). Thus  $\Delta$ TF-IDF considers the biased occurrence of terms with respect to individual classes (sentiment in our case). The  $\Delta$ TF-IDF value  $W_{ij}$  for a term  $t_i$  in a text  $j$  is obtained using:

$$\begin{aligned} W_{i,j} &= \Delta TF - IDF = tf_{i,j} \cdot \left( \log_2 \frac{N^+}{df_i^+} \right) - tf_{i,j} \cdot \left( \log_2 \frac{N^-}{df_i^-} \right) \\ &= tf_{i,j} \cdot \left( \log_2 \frac{N^+}{df_i^+} \frac{df_i^-}{N^-} \right) \end{aligned} \quad (4)$$

where: (i)  $N^+$  is the number of positive texts in the input document collection  $D$  (labelled "aye" with respect to our political opinion mining application), (ii)  $N^-$  is the number of negative texts (labelled "nay"), (iii)  $tf_{i,j}$  is the term frequency for term  $t_i$  in text  $j$ , (iv)  $df_i^+$  is the document frequency for term  $t_i$  with respect to positive texts in the input document collect  $D$  and (v)  $df_i^-$  is the document frequency for term  $t_i$  with respect to negative texts.

However, the  $\Delta$ TF-IDF scheme is directed at sentiment classification of individual texts according to their  $\Delta$ TF-IDF values (Martineau and Finin, 2009). Our research is focused on building domain-specific lexicons and thus a slightly adapted  $\Delta$ TF-IDF weighting scheme is proposed so that term weightings are considered with respect to the entire document collection  $D$  and not per document. We will refer to this scheme as  $\Delta$ TF-IDF'. Thus the  $\Delta$ TF-IDF' value  $W_{i,D}$  for a term  $t_i$  with respect to a document collection  $D$  is obtained using:

$$W_{i,D} = \Delta TF - IDF' = tf_i^+ \cdot \left( \log_2 \frac{N^+}{df_i^+} \right) - tf_i^- \cdot \left( \log_2 \frac{N^-}{df_i^-} \right) \quad (5)$$

where:  $tf_i^+$  is the term frequency with respect to positive texts and  $tf_i^-$  is the term frequency with respect to negative texts. The advantages offered by the proposed  $\Delta TF-IDF'$  scheme are that it can be used to assigning sentiment scores to each term taking into consideration term occurrences in both negative and positive texts.

Thus the  $\Delta TF-IDF'$  scheme is used to determine sentiment scores for each term. On completion of this step, each term in the BOW will comprise an 7-tuple of the form  $\langle t_i, post_i, tf_i^+, tf_i^-, df_i^+, df_i^-, s_i \rangle$ , where  $s_i$  is the sentiment score associated with term  $t_i$ .

#### 6.4 Direct and adaptive generation

As the name implies, using the direct generation technique the desired domain-specific sentiment lexicon is directly generated from the labelled source data processed as described above. Thus the generated *BOW*, which contains the terms and their associated POS tags, sentiment scores and polarities, is converted directly into a domain specific lexicon. Our PoLex political domain specific lexicon was generated in this manner. The PoLex lexicon comprises 170,703 terms such that each term is encoded in the form of a tuple  $\langle t_i, post_i, s_i \rangle$ , thus each term in PoLex is combined with its associated part-of-speech tag, sentiment score and polarity (which is simply the sign of the sentiment score).

In the case of adaptive generation the idea is to use an existing, domain-independent, sentiment lexicon *Lex*<sup>8</sup>, and adapt this to produce a domain-specific lexicon *Lex'*. More specifically the content of *Lex* is copied over to *Lex'*. The adaptation is as follows. Given a term  $t_i$  that is both in  $T$  (the list of all distinct terms derived from the document collection) and *Lex*, if the two associated sentiment scores have different polarities (thus one negative and one positive, or vice versa) we adopt the polarity from the calculated sentiment score (but not the magnitude) with respect to *Lex'*. Terms included in  $T$  but not in *Lex* are simply appended to *Lex'*. Our PoliSentiWordNet political domain specific lexicon was generated in this manner. PoliSentiWordNet comprises 258,353 terms.

<sup>8</sup> SentiWordNet 3.0 was used in the case of the work described here.

## 7 Comparison

This section reports on the comparison and evaluation of the three techniques: straight forward classification, use of generic lexicons and use of specific lexicons. With respect to straight forward classification six machine learning classifiers were considered: Naive Bayes, Support Vector Machine SMO, J48 decision trees learner, JRip rules-based classifier, IBk nearest neighbour classifier and ZeroR (the last as a baseline classifier). The generic lexicon used was SentiWordNet 3.0. The domain specific lexicons used were PoLex and PoliSentiWordNet generated as described above. The comparison was conducted using the House of Commons political debate corpus, described in Section 3, which comprised 2086 concatenated speeches. Recall that the classifiers were used to assign predefined attitude class labels ( $\{positive, negative\}$ ) to each record, while the lexicons were used to assign sentiment scores to the each record which were then used to determine attitude label ( $\{positive, negative\}$ ).

Because the attitude of individual speakers with respect to each debate was known from the way that the speakers eventually voted, the predicted attitude could be compared with the known attitude.<sup>9</sup> Thus we made the assumption that speeches made during the course of a debate reflect how speaker will eventually vote, in other words it is assumed that speakers never “change their mind” during a debate. The metrics used for the comparison were precision, recall, the F-measure and average accuracy. The F-measure (the harmonic mean of precision and recall) combines the precision and recall values and is a good overall measure. The following four equations (6a-6d) show how the metrics are calculated:

$$Precision(P) = \frac{TP}{TP + FP} \quad (6a)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (6b)$$

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (6c)$$

$$F - Measure(F) = \frac{2 \times P \times R}{P + R} \quad (6d)$$

where  $TP$ ,  $TN$ ,  $FN$  and  $TN$  are the True Positive, True Negative, False Negative and True Negative counts respectively<sup>10</sup>.

<sup>9</sup> In the few debates in the data set that were followed by multiple votes, it was assumed that the first vote better represented the speaker’s attitude.

<sup>10</sup> True Positive (TP): Speaker says Aye and sentiment miner says Aye. True Negative (TN): Speaker says Nay and sentiment miner says Nay. False Negative (FN): Speaker says Aye and sentiment miner says Nay. False Positive (FP): Speaker says Nay and sentiment miner says Aye.

The rest of this section is organised as follows. The results obtained using straight forward classification and using lexicons are presented in sub-sections 7.1 and 7.2. The classification based approach also allows for the inclusion of additional information (for example party affiliation) in the feature vector representation, which is precluded when using the lexicon based approach. The results obtained from further experiments conducted using such additional information are presented in Sub-section 7.3.

**Table 4** Evaluation results obtained using the classification based approach to sentiment mining (values generated from confusion matrix data given in Table 5).

Classifier	Precision (P)			Recall (R)			F-Measure (F)			Accuracy (A)
	Aye	Nay	Avg.	Aye	Nay	Avg.	Aye	Nay	Avg.	Avg.
<b>J48</b>	0.628	0.601	0.615	0.719	0.497	0.618	0.671	0.544	0.613	61.751%
<b>JRip</b>	0.581	0.529	0.557	0.685	0.418	0.562	0.629	0.467	0.555	56.238%
<b>SMO</b>	0.602	0.531	0.569	0.604	0.529	0.570	0.603	0.530	0.570	56.963%
<b>NB</b>	0.576	0.493	0.538	0.531	0.538	0.534	0.552	0.515	0.535	53.433%
<b>IBk</b>	0.547	0.500	0.525	0.888	0.132	0.541	0.677	0.209	0.462	54.110%
<b>ZeroR</b>	0.541	0.000	0.293	1.000	0.000	0.541	0.702	0.000	0.380	54.110%
<b>Min</b>	0.541	0.000	0.293	0.531	0.000	0.534	0.552	0.000	0.380	53.433%
<b>Max</b>	0.628	0.601	0.615	1.000	0.538	0.618	0.702	0.544	0.613	61.751%
<b>Average</b>	0.579	0.442	0.516	0.738	0.352	0.561	0.639	0.378	0.519	56.101%
<b>SD</b>	0.033	0.220	0.114	0.176	0.230	0.031	0.056	0.223	0.084	3.089%

**Table 5** Confusion matrix for classification based approach to sentiment mining.

	Class label = Aye						Class label = Nay					
	J48	JRip	SMO	NB	IBk	ZeroR	J48	JRip	SMO	NB	IBk	ZeroR
<b>Speaker votes Aye</b>	805	766	676	594	994	1119	314	353	443	525	125	0
<b>Speaker votes Nay</b>	477	552	447	438	824	949	472	397	502	511	125	0

## 7.1 Results obtained using the classification based approach

Regarding the classification based approach the results are presented in Table 4. The classifiers were trained on a proportion of the data and tested on the remainder using Ten-fold Cross Validation (TCV). Two sets of experiments were conducted. Table 4 shows the overall precision, recall and F-measure (with respect to both the Aye and the Nay classes), and the average accuracy, values obtained. From the table it can be observed that good results were obtained using the J48 classifier generator which outperformed all the other classifiers including the SMO classifier, this was a surprising result as SVMs are usually considered to be well suited to text classification (Joachims, 1998). Reasonable results were also obtained using the JRip classifier. The

worst recorded average F-measure (0.380) and worst recorded average precision (0.293) were obtained using the ZeroR classifier, while the worst recorded average recall (0.538) were obtained using the Naive Bayes classifier. Inspection of Table 4 also indicates that there is no discernible difference with respect to the operation of the first five classifiers with respect to either the Aye or the Nay class (not the case when using lexicons as will become apparent below). Note that the ZeroR classifier has only been included to provide a baseline classifier so as to establish a baseline accuracy. ZeroR is a simple rule-based classifier and that only predicts the majority (most common) class. Table 5 shows the confusion matrix data used to calculate the metrics given in Table 4. With respect to Table 5 the True Positive (TP) counts with respect to each classifier are given in the top-left quadrant, the True Negative (TN) counts in the bottom-right quadrant; the False Positive (FP) and the False Negative (FN) counts are given in the top-right and bottom-left quadrants respectively.

## 7.2 Results obtained using the lexicon based approach

Regarding the lexicon based approaches, the results produced using Poxex, PoliSentiWordNet and the general purpose SentiWordNet 3.0 lexicons, are presented in Table 6 (the associated confusion matrix data is given in Table 7). Note that in each case we also compare the use of the feature vector Term Frequency occurrence count approach with the binary occurrence count approach (the columns labelled TF and Binary respectively). Inspection of the results presented in Table 6 indicates: (i) that there is a small improvement with respect to the average values obtained when using domain specific lexicons and (ii) that both Poxex and PoliSentiWordNet produced similar results. Closer inspection reveals the interesting observation that all the lexicon based techniques worked better with respect to predicting positive (Aye) attitudes than negative (Nay) attitudes. The reason for this, it is argued here, is due to the often overly polite parliamentary jargon used which means that positive sentiment is easier to identify than negative sentiment. Comparing Table 6 with the results previously presented in Table 4 it can be seen that the classification based approach tends to produce a better prediction than the lexicon based approaches, a best classification based average accuracy was achieved using JRip (61.751%) compared to a best lexicon based accuracy using the PoLex domain specific lexicon and binary feature vectors (55.464%).

**Table 6** Evaluation results produced using the lexicon based (generic and domain specific) approaches to sentiment mining (values generated from confusion matrix data given in Table 7).

	PoLex			PoliSentiWordNet			SentiWordNet 3.0		
	Aye	Nay	Avg.	Aye	Nay	Avg.	Aye	Nay	Avg.
<b>Precision</b>	0.554	0.503	0.528	0.554	0.500	0.527	0.554	0.495	0.524
<b>Recall</b>	0.798	0.241	0.520	0.777	0.262	0.520	0.766	0.271	0.518
<b>F-Measure</b>	0.654	0.326	0.490	0.647	0.344	0.496	0.643	0.350	0.497
<b>Avg. Accuracy</b>	54.255%			54.110%			53.894%		

	PoLex			PoliSentiWordNet			SentiWordNet 3.0		
	Aye	Binary	Avg.	Aye	Binary	Avg.	Aye	Binary	Avg.
<b>Precision</b>	0.560	0.534	0.547	0.556	0.522	0.539	0.546	0.502	0.524
<b>Recall</b>	0.831	0.229	0.530	0.831	0.217	0.524	0.909	0.109	0.509
<b>F-Measure</b>	0.669	0.320	0.495	0.666	0.307	0.486	0.682	0.179	0.430
<b>Avg. Accuracy</b>	55.464%			54.937%			54.167%		

**Table 7** Confusion matrix for lexicon based (generic and domain specific) approach to sentiment mining.

	PoLex		PoliSentiWordNet		SentiWordNet 3.0	
	TF	Binary	TF	Binary	TF	Binary
<b>TP</b>	893	930	870	930	857	1016
<b>FN</b>	226	189	249	189	262	103
<b>TN</b>	229	217	249	207	258	104
<b>FP</b>	720	732	700	742	691	845
<b>Total</b>	2068	2068	2068	2068	2068	2068

### 7.3 Additional results obtained using variations of the classification based approach

As noted above the classification based approach allows for additional information to be included in the feature vector representation. The authors thus conducted additional experiments that included party affiliation and debate ID information in the data representation. The intuition being that if a speaker with a particular political affiliation in debate  $N$  voted Aye than other people with the same political affiliation debating within debate  $N$  are also likely to vote Aye. It must be noted, however, that the use of such information would restrict the applicability of the framework to debates for which we know how some debaters have voted (e.g., ruling out debates that are in progress or that are concluded but have not yet reached the voting stage). The results are presented in Table 8. Comparing the results presented in Table 8 with the results presented previously in Table 4 it can be observed that significantly better results were obtained when adding party affiliation and debate ID than when using the speech information on its own (best average accuracy of 85.397% compared to a best average accuracy of 61.751%). Overall best performance was obtained using the J48 classifier. Good results were also obtained using the JRip classifier. The worst recorded average F-measure (0.451) was

obtained using the IBk classifier, while the worst recorded average precision (0.542) and average recall (0.538) were obtained using the Naive Bayes classifier.

**Table 8** Evaluation results obtained using the classification based approach to sentiment mining applied to speeches augmented with party affiliation and debate ID).

Classifier	Precision (P)			Recall (R)			F-Measure (F)			Accuracy (A)
	Aye	Nay	Avg.	Aye	Nay	Avg.	Aye	Nay	Avg.	Avg.
<b>J48</b>	0.865	0.841	0.854	0.865	0.841	0.854	0.865	0.841	0.854	85.397 %
<b>JRip</b>	0.892	0.710	0.809	0.688	0.902	0.786	0.777	0.795	0.785	78.627 %
<b>SMO</b>	0.707	0.655	0.683	0.709	0.653	0.683	0.708	0.654	0.683	68.327 %
<b>NB</b>	0.580	0.497	0.542	0.533	0.545	0.538	0.555	0.520	0.539	53.820 %
<b>IBk</b>	0.551	0.589	0.569	0.945	0.094	0.554	0.696	0.162	0.451	55.416 %
<b>ZeroR</b>	0.541	0.000	0.293	1.000	0.000	0.541	0.702	0.000	0.380	54.110 %
<b>Min</b>	0.541	0.000	0.293	0.533	0.000	0.538	0.555	0.000	0.380	53.820 %
<b>Max</b>	0.892	0.841	0.854	1.000	0.902	0.854	0.865	0.841	0.854	85.397 %
<b>Average</b>	0.689	0.549	0.625	0.790	0.506	0.659	0.717	0.495	0.615	65.949 %
<b>SD</b>	0.158	0.293	0.205	0.177	0.379	0.137	0.102	0.344	0.189	13.732 %

Given the results presented in Table 8 it is interesting to consider whether a better indicator of attitude is simply party affiliation and debate ID on its own. The authors thus constructed a data set comprising only two features, party affiliation and debate ID, and conducted some further classification experiments with this aim in mind. The results are presented in Table 9. From the table it can be observed that best results were again obtained using the J48 classifier generator which outperformed all the other classifiers (average accuracy of 87.089% compared to a best average accuracy of 85.397% obtained using speeches and party affiliation and debate ID number). Good results were also obtained using the JRip classifier, but there were also surprisingly good results obtained using IBk, NB and SMO. Whatever the case the results clearly show that the best indicator of attitude, given a particular debate (identified by a unique ID), is party affiliation and not the content of concatenated speeches made by individuals. In other words, as might be expected, speakers that belong to the same party are likely to vote in the same way. Thus if we wish to predict the likely outcome of a debate while it is in progress we should determine attitude (using our proposed techniques) with respect to groups of speeches belonging to speakers with the same political affiliation. We could do this either by further concatenating the speeches belonging to speakers with the same affiliation and analysing each as one very large “document” or alternatively by using some voting system to produce an aggregated attitude for groups of speakers with the same affiliation.

**Table 9** Evaluation results obtained using a classifier built using only party affiliation and debate ID.

Classifier	Precision (P)			Recall (R)			F-Measure (F)			Accuracy (A)
	Aye	Nay	Avg.	Aye	Nay	Avg.	Aye	Nay	Avg.	Avg.
<b>J48</b>	0.876	0.865	0.871	0.887	0.851	0.871	0.881	0.858	0.871	87.089 %
<b>JRip</b>	0.896	0.722	0.816	0.705	0.903	0.796	0.789	0.802	0.795	79.594 %
<b>SMO</b>	0.840	0.776	0.811	0.799	0.821	0.809	0.819	0.798	0.809	80.899 %
<b>NB</b>	0.781	0.707	0.747	0.735	0.757	0.745	0.757	0.731	0.745	74.468 %
<b>IBk</b>	0.851	0.840	0.846	0.868	0.821	0.846	0.859	0.830	0.846	84.623 %
<b>ZeroR</b>	0.541	0.000	0.293	1.000	0.000	0.541	0.702	0.000	0.380	54.110 %
<b>Min</b>	0.541	0.000	0.293	0.705	0.000	0.541	0.702	0.000	0.380	54.110 %
<b>Max</b>	0.896	0.865	0.871	1.000	0.903	0.871	0.881	0.858	0.871	87.089 %
<b>Average</b>	0.798	0.652	0.731	0.832	0.692	0.768	0.801	0.670	0.741	76.797 %
<b>SD</b>	0.132	0.325	0.218	0.109	0.342	0.119	0.066	0.331	0.182	11.933 %

## 8 Conclusions

In this paper we have compared the operation of three different mechanisms for conducting sentiment mining in the context of political debates with the objective of predicting their outcome. The three different sentiment mining mechanisms considered were: classification based, generic lexicon based and domain specific lexicon based. In the context of the later, two different mechanisms were also considered for generating the desired domain specific lexicons: direct generation and adaptive generation. The comparison was conducted using a collection of 2086 concatenated speeches for 29 different debates extracted from the proceeding of the UK House of Commons. The conducted comparison indicated that classification based sentiment mining outperformed lexicon based sentiment mining. Out of the lexicon based techniques, using domain specific lexicons produced better results than when using generic lexicons. It was also interesting to note that the lexicon based techniques were better at predicting Aye votes than Nay votes. Additional experiments were conducted, with respect to the classification based approach, to determine whether better prediction results can be produced if we include political affiliation and debate number in the representation. The results indicated that this was indeed the case. Further it was confirmed, perhaps unsurprisingly, that individual speakers with the same political affiliation will vote the same way. Although the use of party affiliation is viable only for debates for which the votes of some debaters are known, the result suggests an interesting direction of research. Thus for future work the authors intend to investigate mechanisms whereby their proposed sentiment mining techniques, especially the classification based technique, can be used to determine the attitude to be associated with groups of speakers with the same political affiliation on the grounds that they are likely to vote together; and in this manner attempt to better predict the outcome of debates while they are in progress.

## References

- Amine A, Elberrichi Z, Simonet M (2010) Evaluation of text clustering methods using wordnet. *The International Arab Journal of Information Technology* 7(4):349–357
- Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA)
- Bellegarda J (2010) Part-of-speech tagging by latent analogy. *Selected Topics in Signal Processing, IEEE Journal of* 4(6):985–993, DOI 10.1109/JSTSP.2010.2075970
- Birla VK, Gautam R, Shukla V (2011) Retrieval and creation of domain specific lexicon from noisy text data. In: *Proceedings of ASCNT-2011*, CDAC, Noida, India, pp
- Birnbaum L (1982) Argument molecules: A functional representation of argument structure. In: *AAAI'82*, pp 63–65
- Chim H, Deng X (2008) Efficient phrase-based document similarity for clustering. *IEEE Transactions on Knowledge and Data Engineering* 20(9):1217–1229
- Choi Y, Cardie C (2009) Adapting a polarity lexicon using integer linear programming for domainspecific sentiment classification. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp 590–598
- Dang Y, Zhang Y, Chen H (2010) A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems* 25(4):46–53, DOI 10.1109/MIS.2009.105, URL <http://dx.doi.org/10.1109/MIS.2009.105>
- Demiroz G, Yanikoglu B, Tapucu D, Saygin Y (2012) Learning domain-specific polarity lexicons. In: *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pp 674–679, DOI 10.1109/ICDMW.2012.120
- Denecke K (2009) Are sentiwordnet scores suited for multi-domain sentiment classification? In: *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, pp 1–6, DOI 10.1109/ICDIM.2009.5356764
- Esuli A, Sebastiani F (2006) SentiWordNet: A publicly available lexical resource for opinion mining. In: *Proceedings from the International Conference on Language Resources and Evaluation (LREC)*
- Grijzenhout S, Jijkoun V, Marx M (2010) Opinion mining in dutch hansards. In: *Proceedings of the Workshop From Text to Political Positions*, Free University of Amsterdam
- Hariharan S, Srinivasan R (2008) A comparison of similarity measures for text documents. *Journal of Information Knowledge Management*

- 7(1):1–8
- Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: 10th European Conference on Machine Learning (ECML'98), pp 137–142
- Kennedy A, Inkpen D (2006) Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence* 22(2):110–125, DOI 10.1111/j.1467-8640.2006.00277.x, URL <http://dx.doi.org/10.1111/j.1467-8640.2006.00277.x>
- Kim SM, Hovy E (2004) Determining the sentiment of opinions. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '04, DOI 10.3115/1220355.1220555, URL <http://dx.doi.org/10.3115/1220355.1220555>
- Kuhn A, Ducasse S, Gibra T (2007) Semantic clustering: Identifying topics in source code. *Information and Software Technology* 49(3):230–243
- Lau R, Zhang W, Bruza P, Wong K (2011) Learning domain-specific sentiment lexicons for predicting product sales. In: e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on, pp 131–138, DOI 10.1109/ICEBE.2011.55
- Li H, Sun C, Wan K (2009) Clustering web search results using conceptual grouping. In: Proc. 8th International Conference on Machine Learning and Cybernetics, pp 12–15
- Liu B (2012) *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers
- Martineau J, Finin T (2009) Delta tfidf: An improved feature space for sentiment analysis. In: Proc 3rd International ICWSM Conference, pp 258–261
- Montejo-Raez A, Martínez-Cámara E, Martín-Valdivia M, Ureña-López L (2012) Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In: Proc 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp 3–10
- Ohana B, Tierney B (2009) Sentiment classification of reviews using sentiwordnet. In: Proceedings of the 9th IT & T Conference, Dublin Institute of Technology
- Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Found Trends Inf Retr* 2(1-2):1–135, DOI 10.1561/1500000011, URL <http://dx.doi.org/10.1561/1500000011>
- Poomagal S, Hamsapriya T (2011) K-means for search results clustering using url and tag contents. In: Proc. International Conference on Process Automation, Control and Computing (PACC), pp 1–7
- Porter MA, Mucha PJ, Newman MEJ, Warmbrand CM (2005) A network analysis of committees in the u.s. house of rep-

- representatives. Proceedings of the National Academy of Sciences of the United States of America 102(20):7057–7062, DOI 10.1073/pnas.0500191102, URL <http://www.pnas.org/content/102/20/7057.abstract>, <http://www.pnas.org/content/102/20/7057.full.pdf+html>
- Qiu G, Liu B, Bu J, Chen C (2009) Expanding domain sentiment lexicon through double propagation. In: Proceedings of the 21st international joint conference on Artificial intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'09, pp 1199–1204, URL <http://dl.acm.org/citation.cfm?id=1661445.1661637>
- Ringsquandl M, Petković D (2012) Expanding opinion lexicon with domain specific opinion words using semi-supervised approach. In: In Proceedings of BRACIS – WTI, Curitiba, Brasil, October 2012
- Rissland EL (1999) I yield one minute...: an analysis of the final speeches from the house impeachment hearings. In: Proceedings of the 7th international conference on Artificial intelligence and law, ACM, New York, NY, USA, ICAIL '99, pp 25–35, DOI 10.1145/323706.323712, URL <http://doi.acm.org/10.1145/323706.323712>
- Salah Z, Coenen F, Grossi D (2013a) Extracting debate graphs from parliamentary transcripts: A study directed at UK House of Commons debates. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (ICAIL 2013), Rome, Italy, pp 121–130
- Salah Z, Coenen F, Grossi D (2013b) Generating domain-specific sentiment lexicons for opinion mining. In: Proceedings of the 9th International Conference on Advanced Data Mining and Applications (ADMA 2013), Hangzhou, China
- Stone PJ, Dunphy DC, Smith MS, Ogilvie DM (1966) The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA
- Thelwall M, Buckley K (2013) Topic-based sentiment analysis for the social web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology 64(8):1608–1617, DOI 10.1002/asi.22872, URL <http://dx.doi.org/10.1002/asi.22872>
- Thomas M, Pang B, Lee L (2006) Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Empirical Methods in Natural Language Processing (EMNLP'06), Association for Computational Linguistics, pp 327–335
- Weichselbraun A, Gindl S, Scharl A (2011) Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, New York, NY,

- 
- USA, CIKM '11, pp 1053–1060, DOI 10.1145/2063576.2063729, URL <http://doi.acm.org/10.1145/2063576.2063729>
- Welch S (1985) Are women more liberal than men in the u.s. congress? *Legislative Studies Quarterly* 10(1):125–134
- Wilks Y, Stevenson M (1998) The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering* 4:135–143, DOI null, URL [http://journals.cambridge.org/article\\_S1351324998001946](http://journals.cambridge.org/article_S1351324998001946)
- Zhang J, Peng Q (2012) Constructing chinese domain lexicon with improved entropy formula for sentiment analysis. In: *Information and Automation (ICIA)*, 2012 International Conference on, pp 850–855, DOI 10.1109/ICInfA.2012.6246900