Extracting Debate Graphs from Parliamentary Transcripts: A Study Directed at UK House of Commons Debates

Zaher Salah
Department of Computer Science
University of Liverpool, UK
zsalah@liverpool.ac.uk

Frans Coenen
Department of Computer Science
University of Liverpool, UK
coenen@liverpool.ac.uk

Davide Grossi
Department of Computer Science
University of Liverpool, UK
d.grossi@liverpool.ac.uk

ABSTRACT

The paper proposes a framework—the Debate Graph Extraction (DGE) framework—for extracting debate graphs from transcripts of political debates. The idea is to represent the structure of a debate as a graph with speakers as nodes and "exchanges" as links. Links between nodes are established according to the semantic similarity between the speeches and indicate an alignment of content between them. Nodes are labelled according to the "attitude" (sentiment) of the speakers, positive or negative, using a lexicon based technique founded on SentiWordNet. The attitude of the speakers is then used to label the graph links as being either "supporting" or "opposing". If both speakers have the same attitude (both negative or both positive) the link is labelled as being supporting; otherwise the link is labelled as being opposing. The resulting graphs capture the abstract representation of a debate as two opposing fractions exchanging arguments on related content.

Keywords: Data mining, Sentiment analysis, Debate visualisation.

1. INTRODUCTION

Opinion (Sentiment) mining is concerned with the use of data mining techniques to extract positive and negative feelings, opinions, attitudes and emotions [3], typically embedded within some form of text, concerning some object of interest. This object may be a product, a person, some legislation, a movie, or some kind of happening or topic [12]. Opinion mining is thus directed at the automatic extraction of subjective information embedded in various types of textual data as opposed to objective or factual information. The nature of the textual data used may differ in the degree of subjectivity that is included, thus texts can be described as being emotionally rich or emotionally poor according to the quantity of positive and/or negative subjective words used in the text.

Political opinion mining is a special form of opinion mining concerned with—as the name suggests—the domain of politics. The research described in this paper is directed at extracting argument graphs, describing political debates, from political textual data using sentiment analysis techniques. The graphs are meant to provide an efficacious visualisation of some high-level structure of the debate such as, critically, who talks about similar issues (and to what extent), and who opposes whom (and how strongly).

More specifically, the paper describes the Debate Graph Extraction (DGE) framework whereby political debates can be represented in terms of sets of interconnected nodes, where the nodes represent speakers (debaters) and the links significant interactions between speakers. Interaction between speakers is considered to be significant if there is a high similarity between the (concatenated) speeches made by the individual speakers. Nodes and links are then labelled according to sentiment. Nodes are labelled with the "attitude" of the speaker, positive or negative according to whether they are for or against the motion of the debate. Once the attitude of the speakers (nodes) is known the links may be labelled accordingly. If two nodes connected by a link both have the same attitude label (both positive or both negative) then the link is labelled as being "supporting". If both nodes have different attitude labels (one is positive and the other is negative) the link is labelled as being "opposing". The resulting graphs offer a visualisation of a high-level structure of the debate recording the two opposing fractions. To act as a focus for the work described in this paper the graph generation process was applied to a parliamentary proceedings corpus consisting of verbatim transcripts of debates held within the UK House of Commons. To the best knowledge of the authors, no one has performed such experiments on this kind of data before.

The paper is structured as follows. Section 2 provides some background on lexicon based sentiment analysis and overviews some previous works relevant to this research topic. Section 3 discusses the application domain and our dataset. Section 4 introduces the DGE framework providing also an illustrative example. An evaluation of the DGE framework is discussed in Section 5. Section 6 then provides some conclusions and considers some future extensions of the proposed work.

2. PRELIMINARIES

We start by providing some relevant background for our study: we first introduce sentiment lexicons and then discuss some related work that has applied opinion mining techniques, including sentiment analysis, to the analysis of political debates.

ICAIL '13, June 10 - 14 2013, Rome, Italy

2.1 Sentiment lexicons

Sentiment lexicons are a lexical resource for sentiment classification. They assign sentiment scores and orientations to single words. 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, summating their sentiment scores and then calculating the difference. The result represents the polarity (positive or negative) of the document.

Relatively small size sentiment lexicons which are built manually can be extended starting from a core set of seed positive and negative terms. This set is then expanded 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.

The present paper uses an off-the-shelf sentiment lexicon called SentiWordNet 3.0 [4], which extends the earlier SentiWordNet 1.0 [11].[1] SentiWordNet associates to each synset (i.e., set of synonyms) 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 $Pos(s) + Neg(s) + Obj(s) = 1$. From the point of view of this paper, SentiWordNet has the key advantage, over other available lexicons[2], of covering the largest number of words (SentiWordNet 3.0 covers 117659 words).

2.2 Related work

In sentiment analysis most published research (e.g., [1, 3, 9, 10, 19, 20, 21, 22, 24, 28]) is focused on what might be referred to as "traditional" types of subjective textual data found in blogs, social networks or specialised websites. For example reviews of movies, news articles, commercial products or services. The literature with respect to these traditional approaches is extensive, thus in this section we will limit ourselves to focussing on approaches directly related to work on political sentiment analysis (the topic of interest with respect to the work described in this paper).

In [12] two Opinion 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 based technique that used a subjectivity lexicon. Despite the fact 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. [25] manually surveyed and discussed different types of arguments made in the short (nearly one-minute) speeches given during the last hour of a debate (hearings) held within the United States House of Representatives in December 1998 on the articles of impeachment of President Clinton. The author showed that short speeches' structure are considerably reduced while the longer speeches are more structured and coherent. In [26] 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 belong to the same speaker, or whether they share similar "content") it was demonstrated that this improved the "support" vs. "oppose" classification over the classification of speeches in isolation. The "support"/"oppose" classification and its usefulness in debate visualisation is also argued for in [5] which manually inspected a number of frequent patterns of interaction in argument. In [14] a mechanism was described for visualising the debate structure extracted from meeting notes of the Dutch Parliament in a graph form similar to that proposed later in this paper, where the nodes represented individuals and weighted arrows represented "interruptions". Individual speeches and interruptions were summarised using word clouds. The latter represents work most closely related to that described here. Finally, in [16] the position of political texts on given societal issues is estimated by using scores for relevant words computed from parties' manifestos.

Mark Prisk (Minister of State (Business and Enterprise), Business, Innovation and Skills; Hertford and Stortford, Conservative)
The only chance is that his singing might have been more harmonious than the economic analysis we were given. I did not notice at any point a mention of the enormous - indeed record- debt that we inherited. To be lectured by a party that left the worst Government debt in my lifetime on the prospects of one month-

Iain Wright (Hartlepool, Labour)
That is a long time.

Mark Prisk (Minister of State (Business and Enterprise), Business, Innovation and Skills; Hertford and Stortford, Conservative)
50 years is a long time. When I listened to that, I thought, It is all very well to say that we should be borrowing more and doing this, but it is a shame. It is a particular shame because there is an important issue here that people outside this room are concerned about: how the financial powers will work. It is a shame that there was a pitiful attempt to pretend that there were no borrowing issues, and that tomorrow we could simply borrow because it the money was available. It is a real shame, because there is an important issue at the heart of this.

John Cryer (Leyton and Wanstead, Labour)
Is the **Minister** aware that at the time of the last election, both the deficit and unemployment were falling? They are now both rising. The **Office for Budget Responsibility**, the body set up by the Government, predicts that the deficit will be £180 billion larger at the end of this Parliament than was predicted at the time of the last election.

Mark Prisk (Minister of State (Business and Enterprise), Business, Innovation and Skills; Hertford and Stortford, Conservative)
With respect to the hon. Gentleman, the other thing that we did not hear from the Labour party was mention of the eurozone. According to Labour Members, the only reason businesses are lacking in confidence is entirely to do with the UK's economic policies: there is nothing going on across the channel, it is all calm, they are enjoying their summer holidays and everything is entirely relaxed. When I deal with businesses on a weekly basis, seeking to encourage them to invest in green projects and elsewhere, they constantly refer to the international financial climate, particularly the eurozone, as the reason for hesitating over investing. I had hoped we would have a balanced debate on this issue, but let us address the **amendment** before us, because that is what matters.
On that basis, it will not come as a surprise to the hon. Gentleman that I intend to resist this **amendment** for two main reasons. First, the Government's approach to the bank's future borrowing is the right one. Secondly, legislation is not the right mechanism to govern the bank's borrowing. There are important issues which those wanting to look at the commitment of financial support for this institution are looking to hear about. Before I address these arguments in turn, let me restate that the coalition Government are committed to the **UK Green Investment Bank** growing into a successful, enduring green financial institution.

Figure 1: Fragment of UKHCD Debate 2 as published on the `TheyWorkForYou.com` www site.

[1] SentiWordNet is accessible at `sentiwordnet.isti.cnr.it`.
[2] Cf. [21] for a detailed comparison of SentiWordNet with other popular, though manually built, lexicons.

Debate ID	Result	Num. Speeches	Min. Words	Max. Words	Avg. Words	SD Words	Total Words
D1	D	131	50	4771	513.718	764.184	67297
D2	D	10	61	4707	773.200	1474.421	7732
D3	D	91	50	5514	382.033	1007.695	34765
D4	D	81	51	3696	399.444	676.128	32355
D5	D	84	50	5713	553.417	1271.349	46487
D6	D	38	55	2627	272.000	517.450	10336
D7	D	71	50	5628	534.634	1049.033	37959
D8	D	143	52	5766	375.399	884.215	53682
D9	C	105	50	4542	600.324	763.315	63034
D10	D	94	51	4368	446.670	888.328	41987
D11	C	40	54	5667	649.875	1182.091	25995
D12	C	60	50	3623	418.983	767.552	25139
D13	C	73	51	4205	712.534	888.614	52015
D14	D	72	50	5746	561.792	1145.917	40449
D15	D	66	50	5425	514.591	1029.887	33963
D16	C	117	50	5861	514.889	917.275	60242
D17	D	56	50	4036	451.375	752.994	25277
D18	D	95	52	4333	472.358	806.703	44874
D19	D	71	51	5196	509.127	905.429	36148
D20	D	145	50	5904	364.869	816.706	52906
D21	C	95	50	4667	224.747	537.658	21351
Min.		10	50	2627	224.747	517.450	7732
Max.		145	61	5904	773.200	1474.421	67297
Avg.		82.762	51.333	4856.905	487.904	906.997	38761.571
SD		34.003	2.614	899.859	133.405	232.173	16225.519
Total		1738	1078	101995	10245.978	19046.944	813993

Table 1: UKCHD Dataset Statistics (C = motion Carried, D = motion Defeated, SD = Standard Deviation)

3. DATASET

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 online in XML format (at `TheyWorkForYou.com`) three hours after they take place. Figure 1 shows an extract from a debate transcript taken from the "Enterprise and Regulatory Reform Bill, Clause 4 - The UK Green Investment Bank: financial assistance" debate (Debate number 2 in our dataset). Figure 2 shows the XML mark-up for the same fragment of text. The advantage offered by this collection is that the outcome of the debate is known and thus we can (at least in part) evaluate the veracity of our debate graph constructions so that some confidence can be gained in the technique when it is applied to debate like discussions of all kinds where the result is not known (or not yet known).

The authors extracted the speeches associated with 21 debates from the `TheyWorkForYou.com` www site 6 debates where the motion was carried and 15 where it was defeated. QDAMiner4[3] was used to extract the desired textual information from the XML debate records. For each debate the speeches associated with the same MP were concatenated together. Concatenated speeches with less than 50 words were ignored as it was conjectured that little meaning could be associated with these speeches. The remaining concatenated speeches were collected together to form a single dataset. We will refer to this dataset as the UK House of Commons Debate (UKHCD) dataset. The dataset comprised 1738 concatenated speeches (911 speeches made by speakers who voted Aye and 827 speeches made by speakers who voted Nay) associated with 481 distinct Members of Parliament (MPs). Some statistics concerning this dataset are presented in Table 1. Note that the number of speeches featured in a debate also equates to the number of MPs taking part. The DGE framework incorporates various techniques taken from the domain of document analysis, thus individual concatenated speeches can also be referred to as documents. The average number of words in a concatenated speech was 488. The speaker's vote (Aye or Nay) was used for evaluation purposes (see Section 5).

[3] `http://provalisresearch.com/products/qualitative-data-analysis-software/`

4. THE DGE FRAMEWORK

An overview of the proposed Debate Graph Extraction (DGE) system is presented in Figure 3. The input for the DGE system is a set of concatenated speeches associated with a single debate, the output is a graph representing the structure of the debate. More formally the input to the DGE framework is a set of n concatenated speeches $S = \{s_1, s_2, \ldots, s_n\}$. The output is a graph of the form $G(V, E, L_v, L_E, f_{map})$ where: (i) V is a set of n vertices (one per concatenated speech) such that $V = \{v_1, v_2, \ldots, v_n\}$, (ii) E is a set of m edges such that $E = \{e_1, e_2, \ldots, e_m\}$, (iii) L_V is a set of two vertex labels (positive or negative), (iv) L_E

Term	DF (Aye)	DF (Nay)	DF (Total)	Difference
people	406	338	744	68
cuts	87	38	125	49
change	154	111	265	43
worse	52	17	69	35
simply	101	70	171	31
care	69	39	108	30
confidence	60	31	91	29
recession	42	13	55	29
women	64	36	100	28
military	42	16	58	26
hope	136	120	256	16
existence	15	0	15	15
wonderful	24	10	34	14
deep	21	7	28	14

Term	DF (Aye)	DF (Nay)	DF (Total)	Difference
timetable	23	23	46	0
taxpayer	11	29	40	-18
generous	10	28	38	-18
fully	34	53	87	-19
sustainable	11	33	44	-22
funding	41	64	105	-23
improve	40	63	103	-23
assure	34	59	93	-25
inherited	9	38	47	-29
previous	101	131	232	-30
raises	8	38	46	-30
reduce	38	73	111	-35
encourage	30	72	102	-42
european	59	105	164	-46

Table 2: Document (speech) Frequencies (DFs), with respect to Aye and Nay votes, associated with selected terms occurring in UKHCD dataset.

is a set of two edge labels (supporting or opposing) and (v) f_{map} is some mapping function that maps the vertex and edge labels on to vertices and edges. The DGE framework describes a four phase process: (i) document/data preprocessing, (ii) attitude detection and node labelling, (iii) edge identification and labelling and (iv) debate graph generation. Each of these phases is described in more detail in the following four sub-sections.

4.1 Preprocessing

The input to the DGE framework, as already noted above, is a set of speeches. In terms of text processing each speech can be conceptualised as a document, and in this context each document represents a speaker and contains all the speeches, with respect to a particular debate, of that speaker concatenated together. The pre-processing phase commences with the conversion of all uppercase alphabetic characters to lower case followed by punctuation mark and numeric digit removal.

The next stage is stop word removal. Stop words are words that carry little meaning (such as "and" or "the") to which no particular sentiment can be attached [7, 13, 23], stop words are thus removed from the document set. Given a specific domain there will also be additional words, other than stop words, that occur frequently. In the case of our UKHCD dataset 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 our bespoke stop-word list.

The following stage is to produce a Bag-Of-Words (BOW) representation containing all the remaining words in the document collection (speeches), $BOW = \{t_1, t_2, \ldots, t_{|BOW|}\}$. Each document will then be represented by some subset of the BOW. In fact two BOWs are created, $BOW1$ and $BOW2$. As will be seen, $BOW1$ is used for attitude detection and $BOW2$ is used for edge identification, each is generated in a slightly different manner. The generation of $BOW1$ includes a lemmatisation process while the generation of $BOW2$ includes a stemming process. 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" [13]. For example "compute", "computes", "computer", "computed", "computation" and "computing" will be counted together because they share the common stem "compute". Many mechanisms have been proposed to perform stemming, in the context of the work described in this paper Snowball stemming was used. With respect to sentiment analysis, words like "suffice", "sufficiency", "sufficient" and "sufficiently", which have different Part Of Speech (POS) tags, will typically have different sentiment scores. However, when stemming is applied, these words will be reduced to a single word (stem) and thus share the same sentiment score therefore losing the more appropriate individual sentiment values. An alternative to stemming is lemmatisation which can also be used to reduce the diversity of word forms. 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 [2]. Hence lemmatisation was used with respect to $BOW1$ and stemming with respect to $BOW2$.

The two bags of words are then used to define two feature spaces from which two sets of feature vectors can be generated. The distinction between the two, other than that one incorporated lemmatisation and the other stemming, is that the feature vector elements in the first case hold term frequency counts while the elements in the second case hold term weightings. A document frequency count is simply the number of documents/speeches in which a term appears (a count of one per document). Table 2 shows the document frequency counts for a number of example terms taken from the UKHCD 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

<speech id="uk.org.publicwhip/standing/standing2012-06-19_ENTERPRISE_05-0_2012-06-26a.5.44" speakerid="uk.org.publicwhip/member/40313" speakername="Mark Prisk" time="17:30" url="http://www.publications.parliament.uk/pa/cm201213/cmpublic/enterprise/120626/pm/120626s01.htm#12062715000226" colnum="230">
The only chance is that his singing might have been more harmonious than the economic analysis we were given. I did not notice at any point a mention of the enormous—indeed record—debt that we inherited. To be lectured by a party that left the worst Government debt in my lifetime on the prospects of one month—
</speech>
<speech id="uk.org.publicwhip/standing/standing2012-06-19_ENTERPRISE_05-0_2012-06-26a.5.45" speakerid="uk.org.publicwhip/member/40302" speakername="Iain Wright" time="17:30" url="http://www.publications.parliament.uk/pa/cm201213/cmpublic/enterprise/120626/pm/120626s01.htm#12062715000227" colnum="230">
That is a long time.
</speech>
<speech id="uk.org.publicwhip/standing/standing2012-06-19_ENTERPRISE_05-0_2012-06-26a.5.46" speakerid="uk.org.publicwhip/member/40313" speakername="Mark Prisk" time="17:30" url="http://www.publications.parliament.uk/pa/cm201213/cmpublic/enterprise/120626/pm/120626s01.htm#12062715000228" colnum="230">
50 years is a long time. When I listened to that, I thought, "It is all very well to say that we should be borrowing more and doing this, but it is a shame." It is a particular shame because there is an important issue here that people outside this room are concerned about: how the financial powers will work. It is a shame that there was a pitiful attempt to pretend that there were no borrowing issues, and that tomorrow we could simply borrow because it the money was available. It is a real shame, because there is an important issue at the heart of this.
</speech>
<speech id="uk.org.publicwhip/standing/standing2012-06-19_ENTERPRISE_05-0_2012-06-26a.5.47" speakerid="uk.org.publicwhip/member/40366" speakername="John Cryer" time="17:30" url="http://www.publications.parliament.uk/pa/cm201213/cmpublic/enterprise/120626/pm/120626s01.htm#12062715000229" colnum="231">
Is the Minister aware that at the time of the last election, both the deficit and unemployment were falling? They are now both rising. The Office for Budget Responsibility, the body set up by the Government, predicts that the deficit will be £180 billion larger at the end of this Parliament than was predicted at the time of the last election.
</speech>
<speech id="uk.org.publicwhip/standing/standing2012-06-19_ENTERPRISE_05-0_2012-06-26a.5.48" speakerid="uk.org.publicwhip/member/40313" speakername="Mark Prisk" time="17:30" url="http://www.publications.parliament.uk/pa/cm201213/cmpublic/enterprise/120626/pm/120626s01.htm#12062715000230" colnum="231">
With respect to the hon. Gentleman, the other thing that we did not hear from the Labour party was mention of the eurozone. According to Labour Members, the only reason businesses are lacking in confidence is entirely to do with the UK's economic policies: there is nothing going on across the channel, it is all calm, they are enjoying their summer holidays and everything is entirely relaxed. When I deal with businesses on a weekly basis, seeking to encourage them to invest in green projects and elsewhere, they constantly refer to the international financial climate, particularly the eurozone, as the reason for hesitating over investing. I had hoped we would have a balanced debate on this issue, but let us address the amendment before us, because that is what matters. On that basis, it will not come as a surprise to the hon. Gentleman that I intend to resist this amendment for two main reasons. First, the Government's approach to the bank's future borrowing is the right one. Secondly, legislation is not the right mechanism to govern the bank's borrowing. There are important issues which those wanting to look at the commitment of financial support for this institution are looking to hear about. Before I address these arguments in turn, let me restate that the coalition Government are committed to the UK Green Investment Bank growing into a successful, enduring green financial institution.

Figure 2: The XML mark-up for the UKHCD Debate 2 source presented in Figure 1.

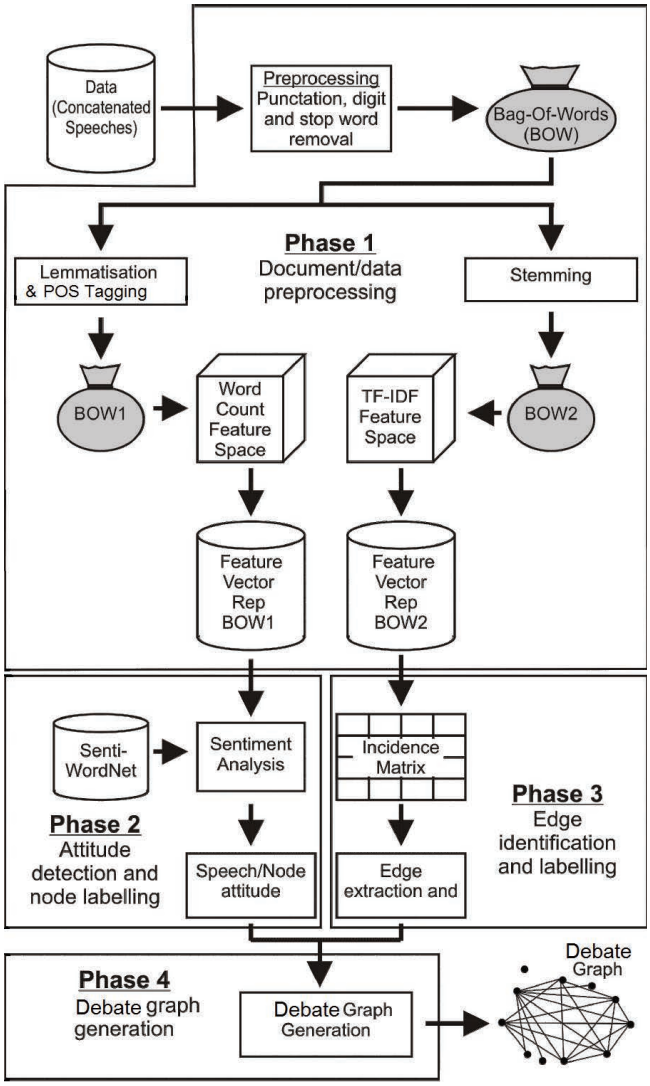

Figure 3: The DGE Framework.

a Nay vote.

The most widely used mechanism for generating term weightings, and that adopted with respect to the DGE framework, is the TF-IDF weighting scheme which aims to "balance out the effect of very rare and very frequent" terms in a vocabulary [15]. TF-IDF also tends to reflect the significance of each term by combining local and global term frequency [17]. TF-IDF can be defined as follows:

$$W_{ij} = \text{TFIDF}(i,j) = tf(i,j).\left(log\frac{N}{df(j)}\right) \qquad (1)$$

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

On completion of the pre-processing phase the input collection of speeches are represented using the vector space model such that each speech can be described by a feature vector. More formally a speech i is represented as a vector $S_i = \{w_{i1}, w_{i2}, \ldots, w_{iz}\}$ where, in the case of BOW1, w_{ij} is the occurrence count of term j in speech i, and in the cases of BOW2 w_{ij} is the TF-IDF value for term j in speech i. It should also be noted that each element in S_i corresponds to a term in either BOW1 or BOW2 as appropriate. We will indicate the list of terms associated with feature vector S_i using the notation $T_i = \{t_{i1}, t_{i2}, \ldots, t_{iz}\}$. Thus we have a set of feature vectors $S = \{S_1, S_2, \ldots, S_z\}$ and a set of term lists $T = \{T_1, T_2, \ldots, T_z\}$ with a one-to-one correspondence between the two.

4.2 Attitude Detection and Node Labelling

From the foregoing, in the case of attitude detection the feature vector weights are simple term frequency counts. Sentiment analysis is then applied to the terms associated with each feature vector to determine node labels (recall that each speech represents a node). The "sentiment" value associated with each term in T_i (the list of terms associated with feature vector S_i) is obtained by "looking up" the term in the SentiWordNet sentiment lexicons.

As mentioned in Section 2 sentiment lexicons assign sentiment scores and orientations to single words. A sentiment score is a numeric value indicating some degree of subjectiv-

ity. 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 result represents the polarity (positive or negative) of the document. SentiWordNet assigns a positive and a negative score (ranging from 0.0 to 1.0) to each synset (semantically similar set of terms) that exists in WordNet so as to generate polarity scores.

In our work, synsets in SentiWordNet have been broken down into single terms in order to produce a list of terms by means of which to retrieve the corresponding score. Terms derived from the same synset are taken to have the same sentiment score. However, if a same term is derived from different synsets then: (i) if the term has different grammatical tagging (POS tag) then word-sense distinction is resolved simply by considering the different POS tags of the term [27] and thus it is split into two distinguished terms; (ii) if the term has the same grammatical tagging in both synsets, then it is treated as a duplicate term and thus the highest sentiment score between the scores of the two synsets is considered.

More formally the sentiment value st_i associated with a speech i is computed using:

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

where SWN is a function that returns the sentiment score for terms from SentiWordNet as single values where each term sentiment score is the summation of the term positivity score (positive value) and the term negativity score (negative value) and thus the value ranges from -1.0 to $+1.0$ and w_{ij} is the frequency of term j in speech i. The attitude of each speaker is then obtained from the total sentiment score st_i. Four types of attitude may be identified: (i) positive (for the motion), (ii) negative (against the motion), (iii) objective (no sentiment scores found) or (iv) neutral attitude (sentiment scores add up to approximately zero). With respect to the evaluation that the authors have carried out to date only positive and negative attitudes have been identified (attitude types 1 and 2). However, should an objective or neutral attitude be discovered the associated node would be excluded from the graph. Algorithm 4.1 describes the node labelling process. The algorithm loops through the input set of speeches, represented in terms of the sets S and T (see end of previous section), a sentiment score for each speech is calculated from lines 7 to 23, the attitude from lines 24 to 38.

Algorithm 4.1 Attitude Identification and Node Labelling

```
 1: INPUT: SentiWordNet dictionary, set of sets of terms
    T^1 ⊂ BOW1, set of feature vectors S^1
 2: OUTPUT: Set of Attitudes labels A = {a_1, a_2, ..., a_z}
 3: PosCount = 0
 4: NegCount = 0
 5: PosScore = 0
 6: NegScore = 0
 7: for all T_i ∈ T^1 do
 8:     for all t_ij ∈ T_i do
 9:         if t_ij ∈ SentiWordNet then
10:             score_ij = SWN(t_ij) × 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 ∧ 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
```

4.3 Link Identification and Labelling

Links between node pairs, as noted above, are established when the speeches associated with two nodes (speakers) are deemed to be similar. There are a number of measures that can be used to determine the similarity between two feature vectors, such as: the Euclidean or Manhatten distance, or the Jaccard measure [18]. For the work described in this paper the cosine similarity measure was adopted because of its wide usage and acceptance. Cosine similarity between a pair of documents d_i and d_j is computed as follows:

$$\begin{aligned} CosSim(d_i, d_j) &= \frac{d_i \times d_j}{|d_i| \times |d_j|} \\ &= \frac{\sum_{k=1}^{k=z} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{k=z} w_{ik}^2 \times \sum_{k=1}^{k=z} w_{jk}^2}} \end{aligned} \quad (3)$$

Cosine similarity is the normalised dot product between two document vectors. Cosine similarity values range between 0 and 1. A value of 1 indicates that the two documents under consideration are identical, and a value 0 means that the two documents are entirely unrelated. With respect to the DGE framework similarities between all document (node) pairs are determined by constructing an affinity matrix. This matrix is then used to determine where links exist between nodes. With respect to the DGE framework a link between two nodes is deemed to exist if the similarity value is greater than the average of all pair-wise similarities. Links are labelled using the terms "support" and "oppose". The label support is applicable if both of the linked nodes have

Algorithm 4.2 Link Identification and Labelling

1: INPUT: Set of feature vectors S^2
2: OUTPUT: Set of Link labels $L = \{a_1, a_2, \ldots, a_z\}$
3: Initialise $z \times z$ affinity matrix *Affinity*
4: **for all** document pairs $\langle s_i, s_{i'} \rangle \in S^2$, $i < i'$ **do**
5: $\quad$ $Affinity_{i,i'} = CosineSimilarity(S_i^2, S_{i'}^2)$
6: **end for**
7: **for all** $Affinity_{i,i'} \in Affinity$ **do**
8: $\quad$ **if** $Affinity_{i,i'} > average\ similarity$ **then**
9: $\quad\quad$ add link to L
10: $\quad$ **end if**
11: **end for**
12: **for all** $l_i \in L$ **do**
13: $\quad$ **if** $l_i.start == l_i.end$ **then**
14: $\quad\quad$ $l_i.label = Support$
15: $\quad$ **else**$[l_i.start \neq l_i.end]$
16: $\quad\quad$ $l_i.label = Opppose$
17: $\quad$ **end if**
18: **end for**

the same attitude, and the label oppose is used if they have different attitudes. The algorithm for determining graph links and their labels is presented in Algorithm 4.2. The input is the set of feature vectors S^2 and the output a list of links L. Each item in L comprises a tuple of the form $\langle start, label, end \rangle$, where *start* and *end* are the start and end node identifiers. To indicate the start or end node, or the label, associated with a particular link l_i the notation $L_i.start$, $l_i.end$ and $L_i.label$ is used. In Algorithm 4.2 the affinity matrix is calculated in lines 3 to 6, this is then processed in lines 7 to 11 to establish the existence of links. The link labels are determined in lines 12 to 17.

4.4 Debate Graph Generation

The final phase of the DGE framework comprises debate graph generation. Graph generation is conducted using the outputs from Algorithms 4.1 and 4.2, and is fairly straightforward. Although any suitable graph drawing package can be used to visualise the generated result the authors used NetDraw[4], a Windows program for visualising social network data [6]. The process supported by the DGE framework can be illustrated using one of the smaller debates from our UKHCD database, for example the "Enterprise and Regulatory Reform Bill, Clause 4 - The UK Green Investment Bank: financial assistance" debate (debate D2). Applying the DGE framework to this debate the graph presented in Figure 4 is generated. With reference to the figure each speaker is represented by a node labelled with a speaker-ID (the official MP ID numbers used in Hansard). A square node indicates a positive attitude and a diamond node a negative attitude. The size of a node reflects the number of links connected to it. The "thickness" of a link between any two speakers reflects the semantic similarity that is calculated by summing the contributions of all terms having non-zero weights (TF-IDF) in both documents covering seemingly related topics (see [8]). Supporting links are indicated by solid links while opposing links are indicated dashed links.

5. EVALUATION

[4]https://sites.google.com/site/netdrawsoftware/home

Debate (C/D)	Num. Nodes	Average Precision	Average Recall	Average Accuracy	Average F-Measure
D1 (D)	131	0.337	0.415	0.397	0.372
D2 (D)	10	0.813	0.700	0.700	0.752
D3 (D)	91	0.430	0.444	0.462	0.437
D4 (D)	81	0.424	0.454	0.481	0.439
D5 (D)	84	0.546	0.527	0.512	0.537
D6 (D)	38	0.346	0.442	0.421	0.388
D7 (D)	71	0.370	0.432	0.437	0.398
D8 (D)	143	0.361	0.442	0.413	0.397
D9 (C)	105	0.370	0.433	0.590	0.399
D10 (D)	94	0.513	0.505	0.479	0.509
D11 (C)	40	0.806	0.672	0.850	0.733
D12 (C)	60	0.615	0.554	0.517	0.583
D13 (C)	73	0.463	0.456	0.849	0.459
D14 (D)	72	0.519	0.510	0.389	0.515
D15 (D)	66	0.503	0.502	0.303	0.502
D16 (C)	117	0.596	0.553	0.607	0.574
D17 (D)	56	0.458	0.479	0.518	0.469
D18 (D)	95	0.432	0.474	0.537	0.452
D19 (D)	71	0.428	0.478	0.507	0.452
D20 (D)	145	0.362	0.454	0.407	0.403
D21 (C)	95	0.607	0.574	0.600	0.590
Min.	10	0.337	0.415	0.303	0.372
Max.	145	0.813	0.700	0.850	0.752
Avg.	82.762	0.490	0.500	0.523	0.493
SD	34.003	0.136	0.076	0.141	0.106

Table 3: Evaluation results for Aye and Nay classification using the UKHCD collection.

One of the challenges of work on debate graph generation is the lack of "ground truth" data. In some cases it is possible to construct such graphs by hand however this still entails subjectivity and requires considerable resources (to the extent that it is not possible to construct significant benchmark data).

To evaluate the DGE framework we compared the attitudes extracted using the sentiWordNet with the known "attitude" of the speaker defined according to whether, at the end of each individual debate they voted "Aye" or "Nay". In doing so we therefore had to assume that the speakers' attitudes during their speeches reflect how the MP is going to vote. As a consequence speakers are taken to never "change their minds" during a debate. We also had to forget, for the purpose of the evaluation, the numerical "intensity" of the attitude computed by Algorithm 4.1.

We have used the standard data mining performance measures: precision (the effectiveness of a system to correctly categorise records as being of a particular class), recall (the effectiveness of a system to distinguish between classes), accuracy (the ratio of correct classification over all classifications) and F-measure (the average of the precision and recall values). The results are presented in tabular form and show the performance of the proposed DGE framework with respect to each debate in our UKHCD collection individually, and the average and Standard Deviation (SD), with respect to the selected measures. The C and D tags included in column one indicate whether the motion was Carried (C) or Defeated (D).

In evaluating the framework we considered its performance with respect to both the classification of "Aye" and "Nay" attitudes (Tables 4 and 5). The evaluation shows that the lexicon-based sentiment analysis technique built in the frame-

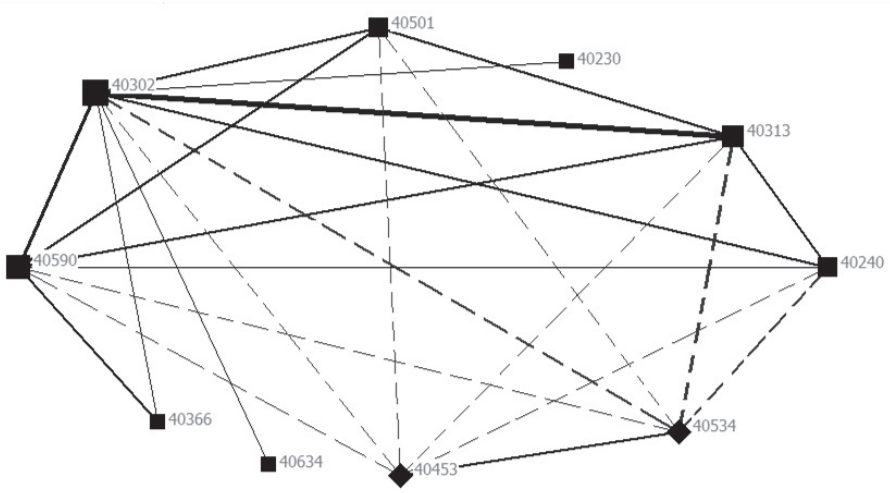

Figure 4: Argument graph generated from UKHCD Debate 2 using the DGE framework.

work performs well with respect to the classification of positive attitudes (Table 4). With respect to Table 4 some high precision, recall and F measure values were obtained. For example with respect to debate D13 the precision, recall and F measure were 0.939, 0.912 and 0.925 respectively, however the D13 debate is unusual in that the ratio of aye votes to nay votes was 68:5. Good results were also obtained for D2, a small debate comprising 10 active participants, with an Aye to Nay ratio of 5:5. The average precision with respect to the Aye class was 0.530; while the average precision with respect to the Nay class was 0.458. Inspection of the tables indicates that the framework exhibits a poorer performance when identifying negative attitudes than when identifying positive attitudes. It can also be observed that better performance on the classification of Aye attitudes is obtained when the class priors are balanced in favour of the "Aye" class. This, we argue, is due to the often overly polite parliamentary jargon that is a feature of House of Commons debates. This issue could be rectified by basing DGE on a dedicated sentiment lexicon for parliamentary language, which could be built using the same data from which we extracted our test set. Accuracy measures, which are independent of class priors, reflect this and the overall average recorded accuracy was 0.523 (Table 3).

Tables 6(a) and (b) show the same data as presented in Table 4 but split into debates where the motion was carried (Table 6(a)) and where the motion was defeated (Table 6(a)). From Tables 6(a) and (b) it can be seen that better precision, recall, accuracy and F-measure values were obtained with respect to debates where the motion was carried than debates where the motion was defeated. Recall that there is often a trade-off between precision and recall although we wish to maximise both.

6. CONCLUSIONS

In this paper we have described the DGE framework for generating debate graphs from transcripts of debates. The objective of the research described was to deploy sentiment analysis techniques for the extraction of debate graphs that will in turn allow for the graphical visualisation of the high-level structure of such debates.

The operation of the framework was illustrated and evaluated using 21 debates taken from the proceedings of the Commons Chamber which are published on-line at `TheyWorkForYou.com` (in XML form). The promising results obtained so far indicate that: (i) it is possible to capture the debate structure representing speakers as nodes using inter-document similarity; (ii) it is possible to use lexicon based opinion mining techniques (such as SentiWordNet) to identify speakers attitudes, although dedicated political lexicons might need to be built to improve overall accuracy.

Future work will initially be directed at the adoption of machine learning techniques (instead of lexicon based ones), more specifically classification techniques, to extract attitude from speeches. The intention is also to increase the size of our UKHCD repository. In the longer term the authors intend to focus on mining of the resulting debate graphs to attempt to predict debate outcomes using the information embedded in their structure.

7. REFERENCES

[1] N. Aleebrahim, M. Fathian, and M. Gholamian. Sentiment classification of online product reviews using product features. In *Proc. 3rd International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA 2011)*, pages 242–245, 2011.

[2] A. Amine, Z. Elberrichi, and M. Simonet. Evaluation of text clustering methods using wordnet. *The International Arab Journal of Information Technology*, 7(4):349–357, 2010.

[3] A. Asmi and T. Ishaya. A framework for automated corpus generation for semantic sentiment analysis. In *Proc. World Congress on Engineering (WCE 2012)*, pages 436–444, 2012.

[4] S. Baccianella, A. Esuli, and F. Sebastiani. 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), 2010.

[5] L. Birnbaum. Argument molecules: A functional

Debate (C/D)	Num. Nodes	Aye Precision	Aye Recall	Aye F-Measure
D1 (D)	131	0.423	0.758	0.543
D2 (D)	10	0.625	1.000	0.769
D3 (D)	91	0.500	0.673	0.574
D4 (D)	81	0.515	0.773	0.618
D5 (D)	84	0.493	0.850	0.624
D6 (D)	38	0.441	0.833	0.577
D7 (D)	71	0.467	0.778	0.583
D8 (D)	143	0.429	0.818	0.563
D9 (C)	105	0.663	0.836	0.739
D10 (D)	94	0.471	0.909	0.620
D11 (C)	40	0.861	0.969	0.912
D12 (C)	60	0.481	0.926	0.633
D13 (C)	73	0.939	0.912	0.925
D14 (D)	72	0.339	0.875	0.488
D15 (D)	66	0.228	0.867	0.361
D16 (C)	117	0.612	0.882	0.723
D17 (D)	56	0.542	0.839	0.658
D18 (D)	95	0.565	0.873	0.686
D19 (D)	71	0.523	0.895	0.660
D20 (D)	145	0.417	0.859	0.561
D21 (C)	95	0.595	0.846	0.698
Min.	10	0.228	0.673	0.361
Max.	145	0.939	1.000	0.925
Avg.	82.762	0.530	0.856	0.644
SD	34.003	0.158	0.073	0.128

Table 4: Evaluation results for Aye classification using the UKHCD collection.

Debate (C/D)	Num. Nodes	Nay Precision	Nay Recall	Nay F-Measure
D1 (D)	131	0.250	0.072	0.112
D2 (D)	10	1.000	0.400	0.571
D3 (D)	91	0.360	0.214	0.269
D4 (D)	81	0.333	0.135	0.192
D5 (D)	84	0.600	0.205	0.305
D6 (D)	38	0.250	0.050	0.083
D7 (D)	71	0.273	0.086	0.130
D8 (D)	143	0.294	0.065	0.106
D9 (C)	105	0.077	0.031	0.044
D10 (D)	94	0.556	0.100	0.169
D11 (C)	40	0.750	0.375	0.500
D12 (C)	60	0.750	0.182	0.293
D13 (C)	73	0.143	0.200	0.167
D14 (D)	72	0.700	0.146	0.241
D15 (D)	66	0.778	0.137	0.233
D16 (C)	117	0.579	0.224	0.324
D17 (D)	56	0.375	0.120	0.182
D18 (D)	95	0.300	0.075	0.120
D19 (D)	71	0.333	0.061	0.103
D20 (D)	145	0.308	0.049	0.085
D21 (C)	95	0.619	0.302	0.406
Min.	10	0.077	0.031	0.044
Max.	145	1.000	0.400	0.571
Avg.	82.762	0.458	0.154	0.221
SD	34.003	0.243	0.105	0.140

Table 5: Evaluation results for Nay classification using the UKHCD collection.

representation of argument structure. In *AAAI'82*, pages 63–65, 1982.

[6] S. Borgatti. *NetDraw Software for Network Visualization*. Lexington, 2002.

[7] H. Chim and X. Deng. Efficient phrase-based document similarity for clustering. *IEEE Transactions on Knowledge and Data Engineerin*, 20(9):1217–1229, 2008.

[8] P. J. Crossno, A. T. Wilson, D. M. Dunlavy, and T. M. Shead. Topicview: Understanding document relationships using latent dirichlet allocation models. In *Proceedings of the IEEE Workshop on Interactive Visual Text Analytics for Decision Making*, 2011.

[9] K. Denecke. Using sentiwordnet for multilingual sentiment analysis. In *Proc 24th IEEE International Conference on Data Engineering Workshop (ICDEW 2008)*, pages 507–512, 2008.

[10] K. Denecke. Are sentiwordnet scores suited for multi-domain sentiment classification? In *Proc. 4th International Conference on Digital Information Management (ICDIM 2009)*, pages 33–38, 2009.

[11] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings from the International Conference on Language Resources and Evaluation (LREC)*, 2006.

[12] S. Grijzenhout, V. Jijkoun, and M. Marx. Opinion mining in dutch hansards. In *Proceedings of the Workshop From Text to Political Positions*. Free University of Amsterdam, 2010.

[13] S. Hariharan and R. Srinivasan. A comparison of similarity measures for text documents. *Journal of Information Knowledge Management*, 7(1):1–8, 2008.

[14] K. Kaptein, M. Marx, and J. Kamps. Who said what to whom? capturing the structure of debates. In *Proc. 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR' 09)*, pages 831–832, 2009.

[15] A. Kuhn, S. Ducasse, and T. Gibra. Semantic clustering: Identifying topics in source code. *Information and Software Technology*, 49(3):230–243, 2007.

[16] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

[17] H. Li, C. Sun, and K. Wan. Clustering web search results using conceptual grouping. In *Proc. 8th International Conference on Machine Learning and Cybernetics*, pages 12–15, 2009.

[18] A. Madylova and S. Oguducu. Comparison of similarity measures for clustering turkish documents. *Intelligent Data Analysis*, pages 815–832, 2009.

[19] J. Martineau and T. Finin. Delta tfidf: An improved feature space for sentiment analysis. In *Proc 3rd International ICWSM Conference*, pages 258–261, 2009.

[20] A. Montejo-Raez, E. Martínez-Cámara, M. Martin-Valdivia, and L. Ureña-López. Random walk weighting over sentiwordnet for sentiment polarity detection on twitter. In *Proc 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 3–10, 2012.

[21] B. Ohana and B. Tierney. Sentiment classification of reviews using sentiwordnet. In *Proceedings of the 9th IT & T Conference*. Dublin Institute of Technology, 2009.

[22] G. Paltoglou and M. Thelwall. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–19, 2012.

[23] S. Poomagal and T. Hamsapriya. K-means for search results clustering using url and tag contents. In *Proc. International Conference on Process Automation, Control and Computing (PACC)*, pages 1–7, 2011.

[24] R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informatics*, 3(2):143–157, 2009.

[25] E. L. Rissland. 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*, ICAIL '99, pages 25–35, New York, NY, USA, 1999. ACM.

[26] M. Thomas, B. Pang, and L. Lee. 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)*, pages 327–335. Association for Computational Linguistics, 2006.

[27] Y. Wilks and M. Stevenson. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4:135–143, 5 1998.

[28] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. HLT/EMNLP-2005*, pages 347–354, 2005.

Debate	Aye Precision	Aye Recall	Aye F-Measure
D9	0.663	0.836	0.739
D11	0.861	0.969	0.912
D12	0.481	0.926	0.633
D13	0.939	0.912	0.925
D16	0.612	0.882	0.723
D21	0.595	0.846	0.698
Min.	0.481	0.836	0.633
Max.	0.939	0.969	0.925
Ave.	0.692	0.895	0.772
SD	0.174	0.050	0.119

(a) Motion carried (aye votes > nay votes)

Debate	Aye Precision	Aye Recall	Aye F-Measure
D1	0.423	0.758	0.543
D2	0.625	1.000	0.769
D3	0.500	0.673	0.574
D4	0.515	0.773	0.618
D5	0.493	0.850	0.624
D6	0.441	0.833	0.577
D7	0.467	0.778	0.583
D8	0.429	0.818	0.563
D10	0.471	0.909	0.620
D14	0.339	0.875	0.488
D15	0.228	0.867	0.361
D17	0.542	0.839	0.658
D18	0.565	0.873	0.686
D19	0.523	0.895	0.660
D20	0.417	0.859	0.561
Min.	0.228	0.673	0.361
Max.	0.625	1.000	0.769
Ave.	0.465	0.840	0.592
SD	0.095	0.076	0.093

(b) Motion defeated (nay votes > aye votes)

Table 6: Evaluation results using the UKHCD collection split into carried (a) and defeated (b) debates.