'Aye' or 'No'? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts

Conference Paper · May 2018

CITATIONS
8

READS
108

2 authors, including:

Gavin Abercrombie
The University of Manchester
9 PUBLICATIONS **33** CITATIONS

SEE PROFILE

‘Aye’ or ‘No’? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts

Gavin Abercrombie and **Riza Batista-Navarro**
School of Computer Science
University of Manchester
Kilburn Building, Manchester M13 9PL
gavin.abercrombie@postgrad.manchester.ac.uk
riza.batista@manchester.ac.uk

Abstract

Transcripts of UK parliamentary debates provide access to the opinions of politicians towards many important topics, but due to the large quantity of textual data and the specialised language used, they are not straightforward for human readers to process. We apply opinion mining methods to these transcripts to classify the sentiment polarity of speakers as being either *positive* or *negative* towards the motions proposed in the debates. We compare classification performance on a novel corpus using both manually annotated sentiment labels and labels derived from the speakers’ votes (‘aye’ or ‘no’). We introduce a two-step classification model, and evaluate the performance of both one- and two-step models, as well as the use of a range of textual and contextual features. Results suggest that textual features are more indicative of manually annotated class labels. Conversely, in addition to boosting performance, contextual metadata features are particularly indicative of vote labels. Use of the two-step debate model results in performance gains and appears to capture some of the complexity of the debate format. Optimum performance on this data is achieved using all features to train a multi-layer neural network, indicating that such models may be most able to exploit the relationships between textual and contextual cues in parliamentary debate speeches.

Keywords: Hansard transcripts, parliamentary debates, sentiment analysis

1. Introduction

In the United Kingdom, transcripts of parliamentary debates (known as *Hansard*) are publicly and freely available. This provides access to a wealth of information concerning the opinions and attitudes of Members of Parliament (MPs) and their parties, towards arguably the most important topics facing society, as well as potential insights into the parliamentary democratic process. However, the large quantity of recorded material in Hansard, combined with the esoteric speaking style and opaque procedural language of Parliament, makes manual retrieval of information from these data a daunting task for the non-expert citizen.

Despite the fact that opinion mining has been one of the most active areas of research in natural language processing (NLP), and a widespread need for political information has been cited as a motivation for the development of opinion mining technologies (Pang and Lee, 2008), automatic analysis of the positions taken by speakers in parliamentary debates has received relatively little attention from researchers.

Sentiment anlysis is the task of automatically identifying the polarity (*positive* or *negative*) of the position taken by the holder of an opinion towards a *target*, such as an organization, a policy, a movement, or a product. We apply sentiment analysis methods to speeches made in the House of Commons of the UK Parliament to classify their sentiment polarity as being either *positive* (in support) or *negative* (in opposition) towards the target of each speech; that is, the *motion* proposed in the debate in question.

Prior work on this task has relied on the use of MPs’ *division votes* as sentiment polarity labels, under the assumption that these votes represent the speakers’ opinions towards the subjects under discussion: votes for ‘Aye’ (that the motion be approved) or ‘No’ (that it be negated) are presumed to indicate *positive* and *negative* sentiment, respectively.

However, as MP voting is to a large extent constrained by party affiliations, with members often under pressure to follow the party whip regardless of their personal opinion (Searing, 1994; Norton, 1997), we perform sentiment analysis experiments on the Hansard Debates with Sentiment Tags (HanDeSeT) corpus, which features manually annotated sentiment labels in addition to those extracted from division votes (Abercrombie and Batista-Navarro, 2018).

In Parliament, the tabled motions under debate, by their nature, either approve of or oppose some piece of legislation or state of affairs, and hence also display sentiment polarity towards those targets. We therefore present a two-stage sentiment analysis model in which first, the sentiment of the motion towards the subject of the debate is determined, before sentiment analysis is carried out on the corresponding speeches.

Our contributions In this paper, we compare the use of speakers’ division votes with manually annotated polarity labels for the evaluation of sentiment analysis systems, and introduce a two-step sentiment analysis model for parliamentary debates in which the sentiment of both *speeches* and *motions* are classified.

For the two-step model, we also propose an alternative method for determining motion sentiment that infers polarity labels from the relationship to the Government of the speakers who introduce the motions

Additionally, we evaluate the use of *n*-gram textual features and a range of contextual features extracted from metadata related to the speakers.

2. Background: UK parliamentary debates

The UK Parliament consists of two chambers: the House of Commons and the House of Lords. The former is the superior legislative chamber, the target of most public and media attention, and the focus of this study.

Each debate in the House of Commons begins with a *motion* proposed by an MP. Following this, MPs may speak, when invited, any number of times during a debate. Each speaking turn may be comprised of a short statement or question, or a longer passage, divided into paragraphs in the transcript.

At any time during a debate, but most typically at the end, a *division* may be called. At this point MPs physically file through one of two *division lobbies* to register their vote—'aye' to support, and 'no' to oppose the motion in question. Labels extracted from the records of these divisions are referred to in this paper as division vote sentiment labels.

3. Related Work

Sentiment analysis has attracted substantial interest in NLP research, where the majority of work focusses on determining people's opinions in product reviews (e.g., Pang et al. (2002), Mukherjee and Bhattacharyya (2012)) and social media posts (e.g., Pak and Paroubek (2010), Rosenthal et al. (2017)).

In the political speech domain, several papers address the application of opinion classification to debates from the United States Congress. For example, Thomas et al. (2006) use a supervised classification model (support vector machine) to determine whether or not individual speech segments support a piece of legislation, using contextual discourse information to obtain enhanced performance, while Burfoot et al. (2011) apply a collective classification approach to Congressional speeches, using the speakers' voting records to obtain sentiment labels. In Europe, Grijzenhout et al. (2010) perform sentiment analysis at the paragraph level on manually labelled Dutch parliamentary transcripts.

For a related but somewhat different task on UK Hansard transcripts, Duthie et al. (2016) present a manually annotated corpus for the detection of speakers' positions, not towards the subject of debate, but rather other members' 'ethos'—which they define as the 'character' of the target, who is another participant in the debate.

For sentiment analysis on this domain, Onyimadu et al. (2013) use a sentiment lexicon to identify opinionated text in House of Commons debates for ternary (*positive*, *negative*, *neutral*) classification at the sentence level, reporting an average accuracy of 43% agreement between a classifier's predictions and the manually applied gold standard labels.

The most similar approach to ours is that of Salah (2014), which compares text classification using machine learning techniques and the use of sentiment lexicons to predict 'speaker attitude' on the concatanated speeches of MPs in the House of Commons, again relying on members' division votes as labels. We challenge the assumption that these votes reflect speaker sentiment by comparing these labels with those of human annotators. We also extend their use of party affiliation information, including other meta information about the debate participants, and examine whether these features are indeed predictive of sentiment as expressed in the speeches, or simply of likely voting outcome.

4. Data: the HanDeSeT corpus

We use the Hansard Debates with Sentiment Tags (HanDeSeT) corpus (Abercrombie and Batista-Navarro, 2018).[1] The corpus consists of 1251 units, each of which is composed of a parliamentary speech of up to five utterances and an associated motion. Content inserted by the Hansard reporters, certain set procedural phrases, and quotations have all been removed from the text.

Each speech has two binary (`1` for *positive* or `0` for *negative*) sentiment polarity labels, produced with different labelling methods:

1. A speaker-vote label extracted from the division associated with the corresponding debate: 'aye' = `1`, 'no' = `0`.
2. A manually annotated gold standard label.

All motions also have been assigned two sentiment labels:

1. A label derived from the party affiliation of the MP who proposes the motion—`1` if they are a member of the governing party or coalition at the time of the debate, `0` otherwise.
2. A manually annotated gold standard label.

In addition, the following metadata is included with each unit: *debate ID*, *speaker party affiliation*, and *motion party affiliation*.

A detailed description of the corpus and annotation process can be found in Abercrombie and Batista-Navarro (2018).

5. Debate speech sentiment models

The motions tabled in these parliamentary debates express either positive or negative sentiment towards a piece of legislation, policy, or state of affairs, and members of the chamber speak either in support of, or in opposition to the motion. For example, a motion may call on members to approve or reject a Bill, Act or Paper, or express approval or condemnation of a policy or situation.

The sentiment polarity of the motion under debate may therefore have a significant effect on the language used by a speaker when either supporting or opposing the motion. For example, for motions that commend the Government, speeches which support the motion are likely to incorporate positive language, while those that oppose the motion will tend to include typically negative language. On the other hand, for motions that oppose Government policy, speeches favourable to the motion are themselves also likely to use typically negative language towards the Government, and unfavourable speeches will conversely use positive language, as in Example 1.[2]

[1] HanDeSeT is available at `https://data.mendeley.com/datasets/xsvp45cbt4`.

[2] For further examples, see Abercrombie and Batista-Navarro (2018).

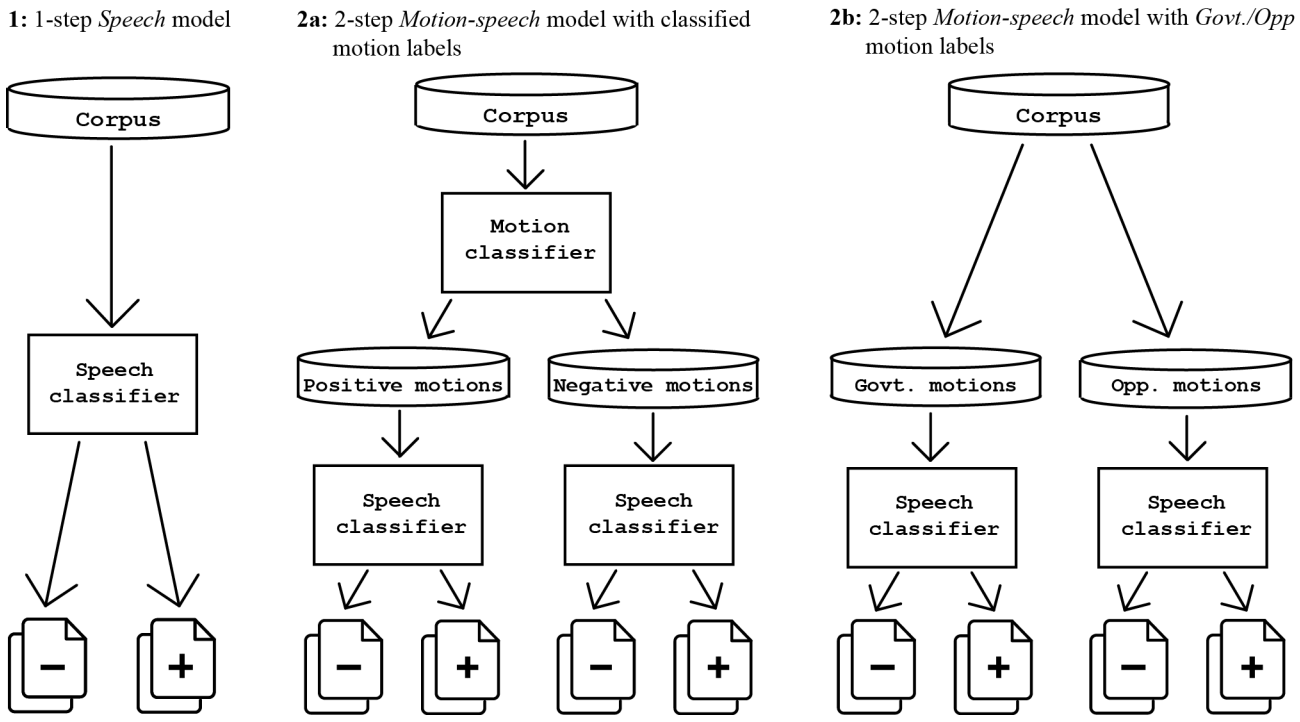

Figure 1: Three classification models for sentiment analysis of parliamentary debates. In model 1, all speeches are classified together, while in models 2a and 2b, speeches given in response to *positive* and *negative* motions are classified separately.

(1) **Motion:** That the Corporation Tax (Instalment Payments) Regulations 1998 (S.I., 1998, No.3175), dated 17th December 1998, a copy of whichwas laid before this House on 17th December, be revoked.

Speech: I do not support the regulations. The Government's rhetoric and practice do not add up. If I may paraphrase a well-respected authority, that which we call a tax riseby any other name would sting as hard, and that would be the effect of the regulations.

In this case, the motion expresses negative sentiment towards a piece of legislation, and the speech (extract) uses negative language to communicate positive sentiment towards the motion.

This 'double negative' effect presents complications for the learning of textual classification features, where lexical features that may be indicative of sentiment can differ in their polarity depending on the sentiment of the motion to which they respond. We therefore propose two models for comparison–as well as two different ways of classifying debate motions (see Figure 1):

1. Model 1: A one-step *Speech sentiment* analysis model, in which all units in the corpus are passed to the classifier simultaneously.

2. Model 2: A two-step *Motion-speech sentiment* analysis model, in which the corpus is first divided into those units with motions expressing positive, and those expressing negative sentiment polarity, before these two groups are classified separately. For this model, we also compare two methods of applying sentiment labels to the motions:

   (a) 2a: Sentiment classification using *n*-gram text features and learned from manually annotated labels.

   (b) 2b: Under the assumption that motions proposed by the Government are positive, and those proposed by other parties are negative, motions are divided by the party affiliation of the MP that proposes them–*positive* if they are a member of the governing party or coalition, *negative* if not.

6. Experiments

We perform experiments to compare sentiment classification performance using combinations of the following:

- Two machine learning models:
  - Support Vector Machines (SVM)–linear support vector classification.
  - Multi-layered Perceptron (MLP)–a neural network with 100 hidden layers, using rectified linear unit (ReLu) activation, L-BFGS optimization and maximum 200 epochs.
- Supervised learning of sentiment polarity classes using both manually annotated labels and division vote labels.
- The two debate models: the one-step *Speech sentiment* model, and the two-step *Motion-speech sentiment* model. For the *Motion-speech* model, we also

compare classification of the motions using *n*-gram textual features with labelling them simply according to the party affiliation of the MP who proposes the motion–*positive* if they are a member of the governing party or coalition, *negative* otherwise.

- The following learning features:
  - Textual features extracted from lowercased, tokenized motions and speeches:
    - *N-grams*: all *uni-*, *bi-*, and *trigrams*, and combinations of these.
  - Contextual metadata features for speech classification:
    - *Speaker party affiliation*. Intuition suggests that a speaker's party membership should be a strong indicator of sentiment towards many topics, and Salah (2014) showed this to be the case, at least as far as correlation with speakers' division votes goes.
    - *Debate ID* number. As there are usually multiple speeches in each debate, and MPs will often express similar sentiments to members of their own party in a particular debate, we also follow Salah (2014) in including this feature to capture possible correlations between MPs' speech and voting behaviour.
    - *Motion party affiliation*. Because MPs are likely to be more or less supportive of a motion depending on who proposes it, we add that Member's party as a further contextual feature.

7. Results & Discussion

We present the results of classification using 10-fold cross-validation. Due to slight imbalances in class labels, F1 scores are reported in addition to accuracy.

For motion classification, the SVM classifier achieves accuracy of 92.1% and an F1 score of 0.921, while the MLP classifier obtains accuracy of 93.0% and an F1 score of 0.931. Considering human agreement rates on this task (Cohen's κ = 0.91[3]), this is probably close to the optimal performance that could be expected.

Many of the features most indicative of positive motion sentiment are related to the practicalities of legislation, reflecting the fact that many of these motions are brought by the Government in an effort to pass law. Many negative motions include structures such as '(this House) believes that/notes that/disagrees with/calls on the Government to...', and this is also reflected in the most discriminating *n*-gram features (see Table 1).

Speech classification performance scores are presented in Table 2. The higest accuracy and F1 scores overall, using both labelling methods, are achieved using all features to train the MLP classifier.

These results provide a number of insights into the relationships between the labelling methods used, the textual and metadata features in the corpus, and the debate models applied.

[3]For more on inter-annotator agreement for this corpus, see Abercrombie and Batista-Navarro (2018).

	Positive	Negative
1	security	notes
2	connection	amend
3	given	believes
4	purposes	calls
5	general	government
6	new	calls government
7	schedule	dated
8	proceedings	eu
9	session	disagrees
10	programme	number

Table 1: Top 10 most discriminating *positive* and *negative* *n*-gram features ranked by SVM training coefficients using manually annotated labels.

7.1. Labelling Methods

Results indicate a correlation between the labelling method used and performance resulting from the use of different feature types for classification. Use of manually annotated labels leads to slightly better performance when only textual features are considered, while with division vote labels, the inclusion (or exclusive use) of meta data leads to considerable gains in performance (see Figure 2).

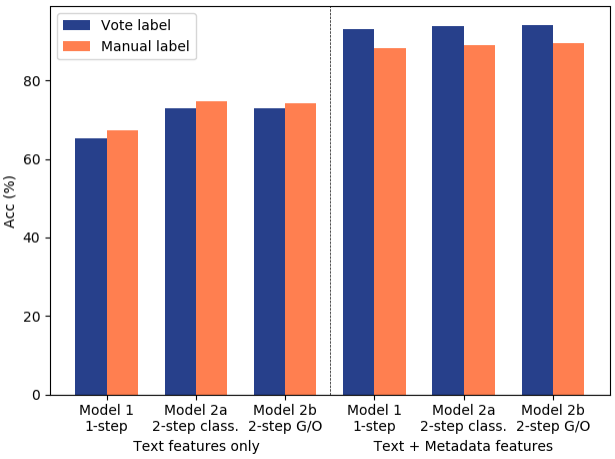

Figure 2: Comparison of manual and division vote labelling methods (using MLP classifier) with contextual features excluded or included.

It therefore appears that information in the text correlates more closely to human understanding of the sentiment expressed in the speech, while contextual information regarding the speakers involved is more indicative of voting intention, with *speaker party affiliation* a particularly strong indicator of this label.

However, while these results support the hypothesis that manual labels are more indicative of speech sentiment, considering the associated costs and the relatively small differences in performance, use of division votes may be the

Debate model	Motion label	Speech label	Features	SVM Acc. (%)	SVM F1 score	MLP Acc. (%)	MLP F1 score
1: *Speech*	n/a	Vote	Text only	64.3	0.699	65.3	0.699
			Text+Party	78.8	0.815	79.2	0.817
			Text+Party+ID	82.7	0.848	87.1	0.888
			Text+Party+ID+Motion	82.6	0.847	93.0	0.938
			Party+ID	83.3	0.853	86.0	0.878
			Party+ID+Motion	83.5	0.854	92.9	0.938
		Manual	Text only	66.7	0.718	67.3	0.713
			Text+Party	76.2	0.791	76.6	0.793
			Text+Party+ID	79.7	0.821	82.4	0.845
			Text+Party+ID+Motion	79.8	0.821	88.2	0.896
			Party+ID	79.9	0.821	82.1	0.842
			Party+ID+Motion	80.0	0.822	88.4	0.897
2a: *Motion-Speech*	Classifier	Vote	Text only	72.9	0.743	72.8	0.739
			Text+Party	83.9	0.835	83.4	0.830
			Text+ID+Party	86.1	0.853	90.7	0.905
			Text+ID+Party+Motion	86.5	0.859	93.9	0.940
			Party+ID	83.3	0.821	91.4	0.915
			ID+Party+Motion	83.2	0.818	93.5	0.935
		Manual	Text only	<u>74.7</u>	0.710	<u>74.6</u>	0.713
			Text+Party	81.0	0.760	81.1	0.772
			Text+Party+ID	83.1	0.794	86.2	0.837
			Text+Party+ID+Motion	82.9	0.790	89.1	0.883
			Party+ID	80.7	0.747	87.1	0.859
			Party+ID+Motion	79.6	0.734	89.0	0.878
2b: *Motion-Speech*	Govt./opp	Vote	Text only	73.1	<u>0.756</u>	72.9	<u>0.748</u>
			Text+Party	85.1	0.853	84.8	0.850
			Text+Party+ID	87.5	0.874	91.7	0.919
			Text+Party+ID+Motion	**87.8**	**0.877**	**94.1**	**0.943**
			Party+ID	82.9	0.820	92.9	0.930
			Party+ID+Motion	84.9	0.848	93.5	0.937
		Manual	Text only	74.3	0.736	74.2	0.736
			Text+Party	72.6	0.799	82.8	0.809
			Text+Party+ID	84.8	0.828	87.4	0.860
			Text+Party+ID+Motion	84.4	0.824	89.6	0.892
			Party+ID	80.8	0.768	88.3	0.876
			Party+ID+Motion	80.6	0.770	89.1	0.885

Table 2: Accuracy and F1 scores for one- and two-step models—the latter using automatically classified motion sentiment labels or *Government/opposition* motion sentiment labels. Results include division vote and manually annotated sentiment labels, and speech sentiment classification is performed using the support vector machine (SVM) and the multi-layer perceptron (MLP) classifiers. The best overall scores for each metric are in bold and best scores using textual *n*-gram features only are underlined.

more pragmatic choice for this task for practical purposes.

7.2. Debate Models

Compared to the one-step *Speech* model, use of the *Motion-speech* models produces improved results for both classifiers under most model-feature configurations. It therefore seems that use of such a two-step model may go some way towards capturing the complex nature of these debates in which positive language can indicate negative sentiment polarity and vice-versa.

Exceptions to this occur when the classifier is trained using contextual metadata features only. Here, as textual features are ignored, the two-step model becomes effectively redundant.

Interestingly, the use in model 2b of labels derived from the relationship of the MP who proposes the motion to the Government (*Government* or *opposition*) is generally as effective as training a classifier on manually annotated labels (model 2a). This suggests that a two-step *Motion-speech* model can be used without the need for costly manual annotations, at least as far as motion sentiment labels are concerned.

7.3. Features

For textual features, the inclusion of bi- and trigrams does not appear to significiantly improve speech classification performance over the use of only unigrams for this task, particularly for the two-step models (see Figure 3).

	1: One-step *Speech* model				2a: Two-step *Motion-speech* model							
	All motions				Positive motions				Negative motions			
	Vote label		Manual label		Vote label		Manual label		Vote label		Manual label	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1	re-search	labour	commu-nity	labour	rural	proper-ly	accept	treaty	tory	labour	week	labour
2	ridicu-lous	money	article	oppor-tunity	hunting	position	impor-tant	pro-gramme mo-tion	place	shadow	tell	shadow
3	street	shadow	ridicu-lous	unfair	press	night	rules	night	commu-nities	suggest-ing	home	snp
4	young people	centres	decis-ion	treaty	open	like	fox	post	impact	snp	young	chilcot
5	work-ing	canna-bis	condi-tions	kent	right	central	increa-ses	prin-ciple	particu-larly	look	yester-day	consult-ation
6	issue	raise	crisis	large	higher	getting	settle-ment	concern	lost	general	public	app-roach
7	higher	leader	people	order	equip-ment	im-posed	pro-gress	in-crease	women	iraq	today	motion
8	cent in-crease	re-quired	early	lowest	dogs	state	poss-ible	floor	conser-vative	centres	welsh	future
9	left	time-table	higher	proper-ly	sub-stan-tial	brought	congra-tulate	head	explain	use	needs	suggest-ing
10	investi-gation	central	young people	coun-cils	area	wales	higher	review	yester-day	benefit	legal	contract
*	0.2	0.2	0.4	0.9	0.2	0.5	0.4	0.2	-0.3	0.5	0.4	0.2

Table 3: Top 10 most discriminating textual n-gram features ranked by coefficients learned by training the SVM classifier. The bottom row of this table (*) shows the total mean sentiment score of the items in each column, as extracted from SentiWordNet 3.0.

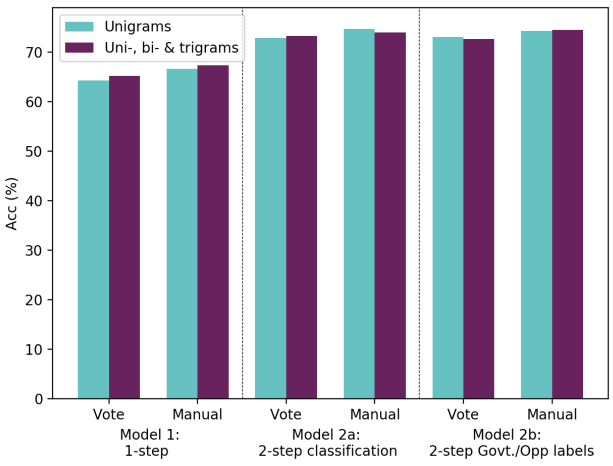

Figure 3: Comparison of MLP classification accuracy using unigram only and uni-, bi-, and trigram textual features. In most configurations, the addition of bi- and trigrams does not notably improve performance over use of unigrams alone.

Ranking of *n*-grams by their SVM training coefficients also reveals that few bigrams and no trigrams feature in the top 10 most discriminating features (see Table 3). Examination of these predictive items underlines the fact that discriminating textual features for this task are not generally those that would be thought of as expressing positive or negative sentiment, even when using the two-step model. Calculating the average polarity of these lexical items (mean score of all entries for each item) according to a sentiment lexicon,[4] we find that 36.7% are neutral, 42.5% positive, and only 16.7% negative. This suggests that MPs tend to follow parliamentary guidelines to practise 'good temper and moderation',[5] avoiding negative language in these debates, whatever point they may be making.

The acquisition of sentiment polarity we see here by objectively neutral language may also be due to the corpus containing a combination of debates on a wide variety of subjects and a relative sparsity of speeches addressing each of these topics. In debates which are skewed towards hav-

[4]SentiWordNet 3.0 (Baccianella, S. and Esuli, A. and Sebastiani, F., 2010), available at `http://sentiwordnet.isti.cnr.it/`.

[5]May (1844) in `https://www.parliament.uk/documents/rules-of-behaviour.pdf`.

ing more speakers either supporting or opposing the motion, topic words can become indicative of one or the other polarity. Hence, in this corpus, generally neutral lexemes such as 'fox' or 'Wales' become indicative of positive and negative sentiment polarity respectively.

While use of contextual metadata features, improves overall performance, in some cases their inclusion leads to incorrect classification. This is prevalent in cases where an MP's sentiment is contrary to that of the majority of other members of their party, or in debates where MPs do not vote along party lines. In such cases, *party affiliation* can be a confounding feature and lead to incorrect classification.

7.4. Classifiers

Using textual features only, there is no significant difference between the performance of the two classifers. However, when contextual metadata features are included, the MLP tends to obtain higher accuracy and F1 scores, suggesting that such neural networks may be better able to exploit the complex relationships between textual and contextual cues in these parliamentary debates.

7.5. Error Analysis

Even using the best performing model-classifier-label-features configurations, some speeches are not classified correctly.

We manually examined the examples for which, using all learning features, and no matter which labels or model were used, the MLP classifier's predicted labels did not match the supervision labels. In the majority of these cases, we observed the following:

1. Speeches were longer than average (μ 218.8 vs. 167.8 words for the whole corpus).

2. Either: speech sentiment labels did not agree with the majority of that speaker's party (19.4% of errors), the speaker's party was split in the debate concerned (11.9%), the speaker was the only member of their party in this debate (22.4%), or the debate featured only that one speech (4.5%).

In the remaining cases, speeches by Conservative MPs were erroneously classified as *negative*, and those of Labour or SNP speakers as *positive*. It therefore appears that the *party affiliation* feature may carry too much weight. While this feature is clearly strongly indicative of speaker sentiment, it can lead the classifier to over-generalise.

For the use of textual features only, we also examined examples in which the best performing (highest accuracy) configuration–the *Motion-speech* model with SVM and manual labels–classified speeches incorrectly. While it is difficult to identify a common thread between all these cases, it appears that on many occasions, these speeches feature speakers addressing off-topic or tangentially related subject matter (see Example 2, in which the speaker talks about a different event than the target of the motion).

(2) **Motion:** That the draft European Union Referendum (Date of Referendum etc.) Regulations 2016, which were laid before this House on 22 February, be approved.

Speech: On suspicious intentions, may I remind the right hon. Gentleman that he campaigned with the Conservative party and the Labour party in Scotland, telling the people of Scotland that if they voted no in the Scottish referendum, they would be guaranteed to remain in the EU? What is his position on that point today?

Even when speeches do contain subjective language directed at the motion, as in Example 3, multiple opinion targets, such as other MPs, parties, and topics, can also be present, complicating the task of sentiment analysis at this level of granularity.

(3) **Speech:** We have always been opposed, and we continue to be opposed, to guillotines. They are wrong in principle and in this case. However, we are realistic and we know that the Government have a majority. We welcome very much the comments and support of the hon. Member for Thurrock...
First, the Bill is unnecessary and should not have been introduced...
As the Government failed to think the matter through and to act, it is unfair that hon. Members should be penalised by lack of time...
Secondly, until a few minutes ago, I was under the impression that the Opposition line was to make their point on the guillotine, but not to divide the House. That will only penalise us, as we will lose another 15 to 20 minutes. I ask the hon. Member for Grantham and Stamford to think.

8. Conclusions

We have evaluated the use of manually annotated labels and division vote labels for sentiment analysis of speeches taken from Hansard UK House of Commons debate transcripts in the HanDeSeT corpus. We have also introduced a new two-step model for debate speech sentiment analysis, and evaluated its performance against the one-step model. We also compared the performance on this task of both SVM and MLP classifiers, and the use of both textual *n*-gram features and contextual metadata features.

Results suggest that while contextual metadata can be highly predictive of their division vote, manually annotated labels more closely reflect speakers' sentiment as expressed in their speeches. However, considering the large overlap between the two sets of labels, for future work or to create larger datasets, manual annotation of these may not be cost-effective.

Our two-step *Motion-speech* model outperforms a simple one-step model in nearly all label-feature-classifier configurations, and therefore seems better able to take account of the complexities inherent in the structure of House of Commons debates, such as double negation. Additionally, we

have found that labelling motions according to the relationship to the Government of the speakers who propose them can approximate the effects of sentiment classification in debate motions, thus avoiding the need for costly manual annotations for this step.

Overall, it seems that sentiment analysis of Hansard transcripts at the speech level does not yield major insights beyond those that could be obtained by merely examining MPs voting records. A more fine-grained analysis may be required to access the opinions expressed in these debates. In future work, we will focus on applying sentiment analysis to the different targets of the speakers' sentiment such as the various topics and subtopics that arise in parliamentary debates.

9. Acknowledgements

The authors would like to thank Loren Hosein, Kieran Lilwall and Anthony Chambers for their contributions to this project, and the anonymous reviewers for their invaluable comments.

10. Bibliographical References

Abercrombie, G. and Batista-Navarro, R. (2018). A sentiment-labelled corpus of Hansard parliamentary debate speeches. In *Proceedings of ParlaCLARIN*. Common Language Resources and Technology Infrastructure (CLARIN).

Burfoot, C., Bird, S., and Baldwin, T. (2011). Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1506–1515. Association for Computational Linguistics.

Duthie, R., Budzynska, K., and Reed, C. (2016). Mining ethos in political debate. In *COMMA*, pages 299–310.

Grijzenhout, S., Jijkoun, V., Marx, M., et al. (2010). Opinion mining in Dutch Hansards. In *Proceedings of the Workshop From Text to Political Positions, Free University of Amsterdam*.

May, T. E. (1844). *A treatise upon the law, privileges, proceedings and usage of Parliament*. C. Knight & Company.

Mukherjee, S. and Bhattacharyya, P. (2012). Feature specific sentiment analysis for product reviews. *Computational Linguistics and Intelligent Text Processing*, pages 475–487.

Norton, P. (1997). Roles and behaviour of British MPs. *The Journal of Legislative Studies*, 3(1):17–31.

Onyimadu, O., Nakata, K., Wilson, T., Macken, D., and Liu, K. (2013). Towards sentiment analysis on parliamentary debates in Hansard. In *Joint International Semantic Technology Conference*, pages 48–50. Springer.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 19–21. European Language Resources Association (ELRA).

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, July.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Salah, Z. (2014). *Machine learning and sentiment analysis approaches for the analysis of Parliamentary debates.* Ph.D. thesis, University of Liverpool, UK.

Searing, D. (1994). *Westminster's world: understanding political roles*. Harvard University Press.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

11. Language Resource References

Abercrombie, G and Batista-Navarro, R. (2018). *HanDeSeT: Hansard Debates with Sentiment Tags*. Mendeley Data.

Baccianella, S. and Esuli, A. and Sebastiani, F. (2010). *SentiWordNet 3.0*. SentiWordNet.