A Sentiment-labelled Corpus of Hansard Parliamentary Debate Speeches

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

Hansard transcripts provide access to Members of Parliament's opinions on many important issues, but are difficult for people to process. Existing corpora for sentiment analysis in Hansard debates rely on speakers' votes as sentiment labels, but these votes are known to be constrained by speakers' party affiliations. We develop an annotation scheme and create a novel corpus designed for use in the evaluation of sentiment analysis systems using automatically and manually applied speech labels. Observing the effects on speech sentiment of differing sentiment polarities in debate motions (proposals), we also apply sentiment labels to these motions. We find that humans are able to reach high agreement in identifying sentiment polarity in these debates, and that manually applied and automatically retrieved class labels differ somewhat, suggesting that speech content does not always reflect the voting behaviour of Members of Parliament.

Keywords: Hansard, UK Parliament, sentiment analysis

1. Introduction

Hansard transcripts of debates from the United Kingdom Parliament provide access to the opinions and attitudes of Members of Parliament (MPs) and their parties towards many important topics facing society. However, the large quantity of recorded material combined with the esoteric speaking style and opaque procedural language used in Parliament makes manual interpretation of information from these data a daunting task for the non-expert citizen.

Sentiment analysis is the task of automatically identifying the polarity (*positive* or *negative*) of the position a person takes towards an entity, such as an organisation, a policy, a movement, a situation, or a product. Automatic detection of MPs' sentiment towards the topics that they discuss in debates has applications in tasks such as information retrieval and question answering, and could allow the public to more easily assess and aggregate the contributions that their elected representatives make in Parliament.

Existing datasets for sentiment analysis of Hansard rely on speakers' votes as sentiment polarity labels (Onyimadu et al., 2013; Salah, 2014). However, it is widely recognised that MPs are to a large extent constrained in their voting behaviour, and often under pressure to vote along party lines irrespective of their personal opinion (Searing, 1994; Norton, 1997). For instance, in Example 1, the speaker appears to be against the motion, yet votes in support of it:

(1) **Motion:** That there shall be an early parliamentary general election.
Speech: Does my right hon. Friend agree that the Prime Minister, in calling this election, has essentially said that she does not have confidence in her own Government to deliver a Brexit deal for Britain? One way in which she could secure my vote and the votes of my hon. Friends is to table a motion of no confidence in her Government, which I would happily vote for.
Vote: 'Aye' (*positive*).

On top of this, MPs may change their mind between speech and vote, and are even known to vote erroneously on occasion.[1] These vote labels may not therefore be accurate reflections of the opinions displayed in the content of MPs' debate speeches, and an alternative form of class labelling may be required for effective sentiment classification using supervised machine learning methods.

Our contribution In this paper, we present Hansard Debates with Sentiment Tags (HanDeSeT), a novel corpus of manually labelled Parliamentary debates for use in the evaluation of automatic Parliamentary speech-level sentiment analysis systems. These consist of proposed *motions* and the associated *speeches* of Members of the House.

2. Related Work

Sentiment analysis has long been one of the most active areas of research in natural language processing (NLP), where attention has been focussed to a large extent on the domains of online reviews (e.g., Pang et al. (2002)) and social media (e.g., Pak and Paroubek (2010)).

For similar tasks in the legislative debate domain, Thomas et al. (2006) use crowdsourced annotations to build a dataset of speech segments from US congressional debates, for which they attempt to automatically determine whether the speakers support or oppose the proposed legislation. Meanwhile, Grijzenhout et al. (2010) create a corpus of Dutch parliamentary debates annotated for *positive* or *negative* 'semantic orientation' at the paragraph level.

In the field of political science, Schwarz et al. (2017) analyse debates from the Swiss parliament, comparing speech content with votes, and find that legislators speak with more freedom than they are able to exercise in their voting behaviour, further motivating our approach.

In the most similar work to ours, Salah (2014) collects a dataset of parliamentary debates comprised of 2,068 speeches in order to perform sentiment analysis on UK Hansard transcripts. Under the assumption that MPs' votes reflect the sentiment of their speeches, these votes are used to label speeches as having *positive* or *negative* polarity.

3. Hansard UK Parliamentary Debates

Hansard transcripts are largely-verbatim records of the speeches made in both chambers of the UK Parliament, in which repetitions and disfluencies are omitted, while supplementary information such as speaker names are added. As the superior legislative body, the House of Commons is

[1] As described by Paul Flynn, MP (Flynn, 2012).

generally of greater interest to the public and media, and is therefore the focus of this study.

3.1. Composition of House of Commons Debates

House of Commons debates consist of these elements:

Motions Debates are initiated with a *motion*—a proposal made by an MP. These motions can be either 'substantive'—requiring the House to support or oppose a policy, piece of legislation, or state of affairs—or 'general'—asking MPs to merely acknowledge that a particular topic has been considered by the House, regardless of their opinions towards it.[2]

Speeches When invited by the *Speaker* (the presiding officer of the chamber), other MPs may respond to the motion, one or more times. Each speaking turn may be comprised of a short statement or question, or a longer passage, which is divided into paragraphs in the transcript.

Divisions At any time (typically at the end of the debate) the Speaker may call a *division*, whereby MPs vote by physically moving to either the 'Aye' or 'No' lobby of the chamber. There may be more than one vote on each motion.[3]

3.2. Semantic Structure of House of Commons Debates

During data collection and initial experiments, we observed certain characteristics of the stucture of these debates which are likely to have a bearing on the sentiment detection task:

Motion sentiment Sentiment polarity is present in both debate speeches *and* motions. In proposing a motion, an MP expresses sentiment towards the policy, piece of legislation, or state of affairs in question.

Double negative effect The language used to express *positive* or *negative* speech sentiment is radically altered depending on the sentiment polarity of the motion. A sort of double negative effect is created, whereby speakers may use typically negative language to demonstrate positive sentiment and vice versa.

For example, if a motion praises the actions of the Government, speeches in support of the motion will likely contain positive language, while those opposing it will be characterised by negative language. If, however, a motion condemns Government policy, supporting speeches are also likely to contain negative language, and opposing ones positive language, as in Example 2:

(2) **Motion:** That an humble Address be presented to Her Majesty, praying that the Local Authorities (England) Regulations 2000 be annulled.

Speech: ... there are deep reservations in the county about all the proposals. I am particularly alarmed about the impact of key decisions. An enormous electoral ward such as Bowbrook or Inkberrow, where huge decisions could be taken affecting communities, will not be subject to openness under the proposals. Why are huge electoral divisions excluded in that monstrous way?

Based on these observations, Abercrombie and Batista-Navarro (2018b) propose a two-stage sentiment analysis model, in which opinions expressed in both motions and speeches are analysed. For this reason, we include manually annotated labels for motions as well as for speeches. Noting that speeches are often made in either attack or defence of the Government's actions, we also include a motion sentiment label derived from the party affiliation of the MP who proposes the motion: *positive* if they are a member of the governing party or coalition, *negative* if not.

4. Corpus Construction

We create and make available a corpus of labelled debates for speech-level sentiment analysis on Hansard debate transcripts from the House of Commons of the UK Parliament.

4.1. Data Collection

Debate transcripts from 1935 onwards are available in XML format on the parliamentary monitoring website TheyWorkForYou.com.[4] In order to obtain a sufficient quantity of speeches for which there are associated division votes, we downloaded the records of all debates in the House of Commons from May 1997 to July 2017.

Each file contains transcipts of a number of debates. We selected all debates under 'major-heading' elements in the XML files—debates which often culminate in *divisions*, or votes. We retained only debates that contain a motion and precisely one *division*, under the assumption that each member's vote represents their sentiment towards the motion under debate. We included only debates with substantive (rather than general) motions, as, by their nature, these demand polarised stances to be taken by MPs.

4.2. Data Processing

Parliamentery speeches incorporate much set, formulaic discourse related to the operational procedures of the chamber, which we automatically removed as it does not concern the motion or the speakers' opinions towards it. These include speech segments such as those used to thank the *Speaker*, or to cede the floor, as well as descriptions of activity in the chamber inserted into the transcripts by the reporters, for example showing that a member rose from their seat or indicated assent by nodding.[5] Additionally, we removed all utterances produced after a division is made, as these are generally procedural matters related to the running of Parliament and/or off-topic.

As in Salah (2014), we consider a member's *speech* to be the concatanated content of *all* their *utterances* (individual speech segments or paragraphs). For comparison of manual and vote labelling methods, we retained all speeches made by MPs who appear in the division of the given debate along with a record of their vote. We omit speeches made by the member of the assembly that proposes the motion, as, by definition, they speak in support of the proposal.

[2] See www.parliament.uk/about/how/business/debates.

[3] For example, several clauses or ammendments to a Bill or Paper discussed in a motion may be voted on individually.

[4] https://www.theyworkforyou.com/pwdata/scrapedxml/debates/

[5] We automatically remove the following procedural language: names of MPs mentioned in speeches (inserted by the reporters), utterances solely concerned with 'giving way' or making interventions, utterances concerning *points of order*.

Also following Salah (2014), we removed speeches totalling fewer than 50 words. In order to facilitate manual labelling, we restrict the quantity of text to be read by human annotators by including only those speeches comprised of five utterances or fewer. Finally, quotations within speeches were removed (and replaced with the word 'QUOTATION'), as these can reflect opposing or different points of view to those of the speaker, and represent confounding features for automatic sentiment classification.

5. Annotation

Annotation guidelines were developed in a two-round cycle using a randomly selected subsection (20%) of the corpus and three annotators—all L1 English speakers, university graduates, UK citizens, and self-reporting as being familiar with British politics and the UK parliament. The principal annotator (*annotator 1*, an author of this paper) then produced the gold standard labels for the complete corpus following the final version of these guidelines.

5.1. Development of Annotation Guidelines

Manual sentiment labelling was carried out on the corpus subsection (250 speech units) in two rounds of the following cycle:

1. Annotation guidelines produced/updated.
2. Two annotators labelled the corpus subsection.
3. Inter-annotator agreement calculated and disagreement analysis performed.

Finally, the principal annotator labelled the full corpus.

5.2. Annotation Guidelines

Following the final annotation guidelines,[6] the job of the annotator can be summarised as follows:

For each *unit* (motion plus speech) in the dataset, the annotator reads the motion carefully, makes a decision on its sentiment polarity towards the subject of the debate, and assigns it the corresponding label: '`1`' for *positive*, '`0`' for *negative*. The annotator then reads the speaker's utterances, considering their overall sentiment polarity, and assigns a label for the sentiment polarity of the speech in question towards the motion (again '`1`' or '`0`').

6. Analysis of the Annotations

To assess the validity of the manually applied labels, we calculated Cohen's kappa (κ) after each round of annotations. We then performed a systematic manual analysis of cases on which the annotators disagreed, identified measures that could be taken to improve agreement, and updated the annotation guidelines accordingly.

Annotation guidelines used	Motion κ	Speech κ
Version 1 (annotators 1 & 2)	0.56	0.57
Version 2 (annotators 1 & 3)	0.91	0.85

Table 1: Inter-annotator agreement (Cohen's kappa) for motion and speech sentiment polarity labels following the first and second versions of the annotation guidelines.

Identified causes of disagreement are presented in Table 2.

Cause of disagreement	Cases (%)
Motions	
Motion calls for action (*positive*), but opposes the target (*negative*)	85.0
Annotator error: same motion labelled differently in other examples	5.0
Annotator error: possible missed negation in motion	5.0
Possible misinterpretation: motion sentiment is against previous, not current Govt.	5.0
Speeches	
Off-topic speech content	16.7
Contextual information required	13.0
Procedural (i.e., long, detailed) motion	13.0
Motion IA disagreement	9.3

Table 2: Causes of annotator disagreement for round 1.

Round one Inter-annotator agreement on the first round of annotation was found to be 'moderate'[7] for both motions and speeches. This was poorer than expected, as intuitively the task seemed relatively straightforward, particularly for labelling of motions, which by definition in these substantive debates are proposed in favour of, or against something.

Round two To address the issues raised by this analysis, we updated the annotation guidelines, clarifying the instructions and adding further example cases. In particular, we defined a protocol for handling motions which call on the Government for action, but which can be seen as attacking its position. These are common in the corpus and had accounted for 85% of annotator disagreement on motion sentiment. We also provided the annotators with more contextual information, by including the MPs' party affiliation. This process resulted in 'very good' agreement on both motions and speeches for the second round of annotations, a considerable improvement on the first round. Given sufficiently clear instructions, humans appear to be capable of high levels of agreement in recognising sentiment polarity in parliamentary debates. As anticipated, sentiment identification in motions seems to be particularly straightforward. We manually analysed cases of disagreement in the second round of annotation, and found that the only two cases of disagreement over motion sentiment were probably caused by error or misinterpretation by one of the annotators. The same can be said for many cases of speech sentiment disagreement, although some were identified as being either off-topic or highly ambiguous, as in Example 3:

(3) **Motion:** That this House believes that the UK needs to stay in the EU because it offers the best framework for trade, manufacturing, employment rights and cooperation to meet the challenges the UK faces in the world in the twenty-first century; and notes that tens of billions of pounds worth of investment and millions of jobs are linked to the UK's membership of the EU, the biggest market in the world.

[6] Available in Abercrombie and Batista-Navarro (2018a).

[7] As described by Landis and Koch (1977).

> **Speech:** My hon. Friend is making a powerful speech and makes an important point about patriotism. Does he agree that key to Britain's national security is our economic security, and at a time when we are still borrowing as a nation more than the entire defence budget we need every single penny of public revenue to ensure our economy is strong, our finances are strong and our country is strong?

Here, without access to information about the speaker's views on a range of issues (e.g., the UK's membership of the EU), the speaker's sentiment towards the motion is likely to seem ambiguous. The presence of such speeches in House of Commons debates makes it unlikely that 100% agreement could be achieved on this task without further contextual clues.

7. Corpus Description

The corpus is available at `https://data.mendeley.com/datasets/xsvp45cbt4/`. It consists of 1251 motion-speech units taken from 129 separate debates, with each unit comprising a parliamentary speech of up to five utterances and an associated motion. Debates comprise between one and 30 speeches, and speeches range in length from 31 to 1049 words, with a mean of 167.8 words. The debates cover a two decade period from 1997 to 2017 and a wide range of topics from domestic and foreign affairs to procedural matters concerning the running of the House.

Each motion has both a manually applied and a *Goverment/opposition* sentiment label and each speech also has two sentiment polarity labels, produced with different labelling methods for comparison: (1) A speaker-vote label extracted from the division associated with the corresponding debate; and: (2) A manually assigned label.

In addition, the following metadata is included with each unit: *debate id*, *speaker party affiliation*, *motion party affiliation*, *speaker name*, and *speaker rebellion rate*.[8]

Manually applied motion labels are approximately evenly balanced; the other labels are slightly skewed towards the positive class (See Table 3).

Target	Label type	Positive	Negative
Motion	Govt./opp.	71 (55.0%)	58 (45.0%)
	Manual	67 (51.9%)	62 (48.1%)
Speech	Vote	713 (57.0%)	537 (43.0%)
	Manual	702 (56.5%)	544 (43.5%)

Table 3: Occurrences of sentiment labels in the corpus.

Concurrence between the vote labels and manually annotated labels is 92.8%. This indicates that, while the majority of speeches reflect the voting behaviour of the speaker, a significant number do not, and that division votes may not therefore be reliable sentiment polarity labels.

In general, MPs both speak and vote along party lines. Of the seven parties that have had more than one sitting MP at a time, the SNP, and the three smaller parties (DUP, UUP, Green) always vote as a block and are assigned the same manual sentiment label. The major UK-wide parties exhibit rather more rebellion, for speech sentiment (Lab: 8.2%, Con: 6.6%, LD: 2.8%) and vote (Lab: 4.2%, Con: 1.1%, LD: 0.0%).

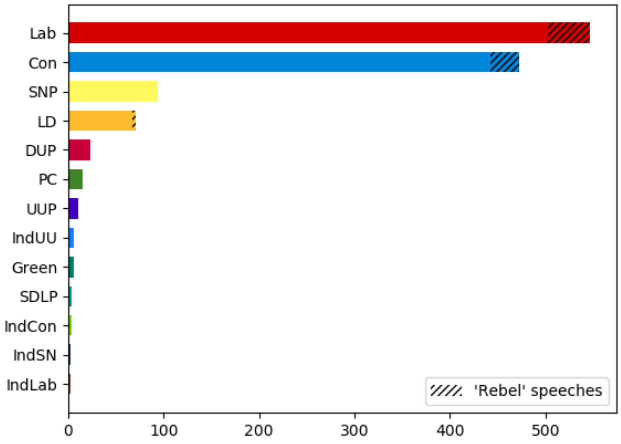

Figure 1: Number of speeches in the corpus by party, and number of 'rebel' speeches in which the manual sentiment label opposes the majority of the speaker's own party.

Examining these 'rebel' speeches, we find that they tend to occur in debates that concern (a) topics of local interest, such as local government finance, in which MPs' loyalties may be divided between party and constituency, (b) matters of conscience, such as stem cell research, or (c) issues that are known to divide parties, such as membership of the EU. In several speeches, a speaker states explicitly that they will vote one way, only to vote for the opposing side, confirming the unreliabilty of votes as sentiment labels.

8. Conclusion

This paper presents a corpus of parliamentary debates from the UK House of Commons, manually annotated and vote-labelled for sentiment at the speech level and with additional sentiment labels applied to debate motions. In order to create this corpus, we developed a set of annotation guidelines, and demonstrated that, using these instructions, agreement on this sentiment identification task can be relatively straightforward for humans, although some ambiguous cases remain challenging. We obtained satisfactory inter-annotator agreement scores, which validate the corpus, and created gold standard labels for use in the evaluation of automatic sentiment analysis systems.

While the majority of manually annotated and automatically applied labels in the corpus agree, a number differ. This indicates that MPs may be freer to express personal opinion in their speeches than in their voting behaviour, and has implications for automatic sentiment analysis, where division votes are perhaps not the best labels for this task.

9. Acknowledgements

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

[8] Rate of MPs' votes against the majority of their party in the current parliament, extracted from http://www.publicwhip.org.uk/.

10. Bibliographical References

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

Abercrombie, G. and Batista-Navarro, R. (2018b). 'Aye' or 'no'? Speech-level sentiment analysis on Hansard UK Parliamentary debate transcripts. In *Language Resources and Evaluation Conference*. LREC.

Flynn, P. (2012). *How to be an MP*. Biteback.

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*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

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.

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.

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

Schwarz, D., Traber, D., and Benoit, K. (2017). Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods*, 5(2):379–396.

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.