Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5073–5078
Marseille, 11–16 May 2020

ParlVote: A Corpus for Sentiment Analysis of Political Debates

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

Abstract

Debate transcripts from the UK Parliament contain information about the positions taken by politicians towards important topics, but are difficult for people to process manually. While sentiment analysis of debate speeches could facilitate understanding of the speakers' stated opinions, datasets currently available for this task are small when compared to the benchmark corpora in other domains. We present *ParlVote*, a new, larger corpus of parliamentary debate speeches for use in the evaluation of sentiment analysis systems for the political domain. We also perform a number of initial experiments on this dataset, testing a variety of approaches to the classification of sentiment polarity in debate speeches. These include a linear classifier as well as a neural network trained using a transformer word embedding model (BERT), and fine-tuned on the parliamentary speeches. We find that in many scenarios, a linear classifier trained on a bag-of-words text representation achieves the best results. However, with the largest dataset, the transformer-based model combined with a neural classifier provides the best performance. We suggest that further experimentation with classification models and observations of the debate content and structure are required, and that there remains much room for improvement in parliamentary sentiment analysis.

Keywords: sentiment analysis, parliamentary debates, Hansard

1. Introduction

Transcripts of debates held in the United Kingdom (UK) Parliament are publicly and freely available and provide access to the opinions and attitudes of Members of Parliament (MPs) and the political parties they represent towards the most important topics facing society and its citizens, as well as potential insights into the parliamentary democratic process. As a result, they are of interest to the politicians themselves, the media, social scientists and historians, and any members of the public who wish to scrutinise the activities of their elected representatives. However, due to the large quantity and complexity of the material, processing the transcripts and analysing the positions taken by the speakers can be a difficult and overwhelming task for humans (Salah, 2014; Thomas et al., 2006).

There has therefore been considerable interest in applying automatic sentiment analysis methods to political debates from legislatures such as the US Congress (Bhatia and P, 2018; Ji and Smith, 2017) and the UK Parliament (Bhavan et al., 2019; Salah et al., 2013). However, in comparison with those currently used in other domains, such as product reviews or social media, which can run into the hundreds of thousands of labelled examples, the datasets currently available for this task are relatively small (see Table 1). As current neural network and embedding-based state-of-the-art sentiment analysis methods tend to benefit from significantly larger datasets, there is a need to develop more extensive corpora for this task.

Our contributions We compile and make available for the research community a large labelled corpus (34,010 examples) of English language UK parliamentary debate speeches labelled at the speech level for use in the evaluation of supervised speech-level sentiment classification systems.

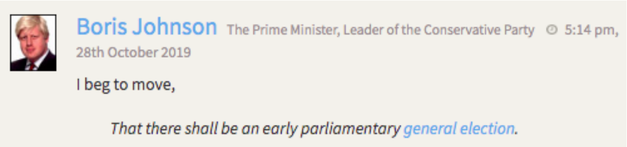
We present the results of initial experiments in which we apply a range of linear and neural machine learning methods and different approaches to text representation to the classification of speeches from a subset of this corpus. We also investigate the effects of increases in dataset size for this task by testing these systems on various subsets of the corpus, and experiment with limiting the length of the input texts.

Dataset	Authors & year	Size
ConVote	Thomas et al. (2006)	3,857
HanDeSet	Abercrombie & Batista-Navarro (2018b)	1,251
ParlVote	This paper	34,010

Table 1: Size in number of example speeches of publicly available datasets for supervised speech-level sentiment analysis of legislative debate transcripts.

2. Background

The UK Parliament consists of two main debating chambers: the House of Commons and the House of Lords. The former is the superior legislative chamber, the target of the majority of public and media attention, and the focus of this work. Each debate in the House of Commons begins with a *motion*—a proposal made by an MP. Motions always begin with the words '*I beg to move, ...*', which are followed by one or more statements (see Figure 1 for an example).


Boris Johnson The Prime Minister, Leader of the Conservative Party 5:14 pm, 28th October 2019

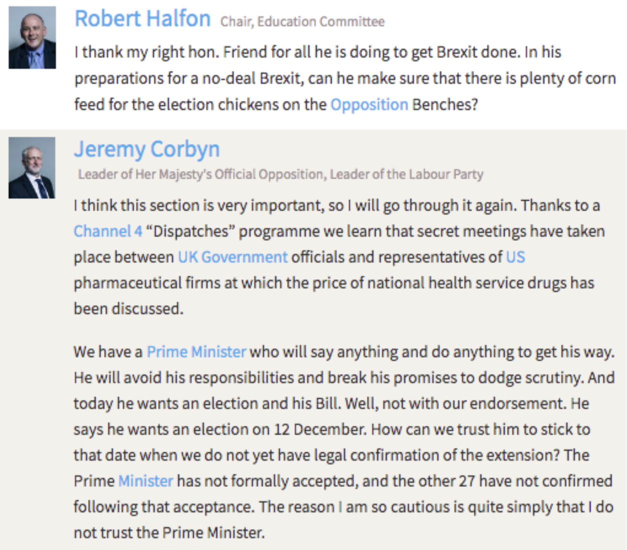
I beg to move,

That there shall be an early parliamentary general election.

Figure 1: Example of a debate *motion* from the corpus as presented on the TheyWorkForYou website.

During the subsequent debate, other MPs may propose one or more counter motions, to amend the wording of the original.

In reponse to the motion, MPs may speak, when invited by the *Speaker* (chief officer of the House), 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 (see Figure 2 for examples). As in previous work, we refer to each of these speaking turns as an *utterance* and the concatenation of a speaker's utterances in a given debate as a *speech*.



Robert Halfon Chair, Education Committee

I thank my right hon. Friend for all he is doing to get Brexit done. In his preparations for a no-deal Brexit, can he make sure that there is plenty of corn feed for the election chickens on the Opposition Benches?

Jeremy Corbyn

Leader of Her Majesty's Official Opposition, Leader of the Labour Party

I think this section is very important, so I will go through it again. Thanks to a Channel 4 "Dispatches" programme we learn that secret meetings have taken place between UK Government officials and representatives of US pharmaceutical firms at which the price of national health service drugs has been discussed.

We have a Prime Minister who will say anything and do anything to get his way. He will avoid his responsibilities and break his promises to dodge scrutiny. And today he wants an election and his Bill. Well, not with our endorsement. He says he wants an election on 12 December. How can we trust him to stick to that date when we do not yet have legal confirmation of the extension? The Prime Minister has not formally accepted, and the other 27 have not confirmed following that acceptance. The reason I am so cautious is quite simply that I do not trust the Prime Minister.

Figure 2: Examples from TheyWorkForYou of utterances made in response to the motion in Figure 1 by speakers who voted *aye* and *no* respectively.

Sentiment analysis is often framed as the task of automatically identifying the polarity (usually *positive* or *negative*) of the position taken by the holder of an opinion towards a target, such as an organization, a movement, or a product (Liu, 2012). In this paper, we consider the target of sentiment for each speech to be the motion preceding it.

At any time during a debate, but most typically at its 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. Because these labels have been found to closely reflect the sentiment perceived by human readers of the transcripts (Abercrombie and Batista-Navarro, 2018a), like the majority of previous work in this domain (Salah, 2014; Thomas et al., 2006), we use the records of these votes to obtain sentiment labels for the corresponding speeches of the MPs.

3. Related Work

For some time, sentiment analysis has been one of the most active research areas in the field of natural language processing (NLP), where the majority of efforts have focussed on the domains of product reviews and social media.

Early work on sentiment analysis of political debates began with Thomas et al. (2006), who constructed the *ConVote* corpus of United States congressional floor debates labelled with speaker roll-call votes. This has been widely used as a benchmark dataset for binary sentiment polarity classification (Balahur et al., 2009; Burfoot, 2008; Burford et al., 2015; Ji and Smith, 2017; Yessenalina et al., 2010; Yogatama et al., 2015), but is limited to less than four thousand labelled examples.

For UK parliamentary transcripts, the *HanDeSet* corpus (Abercrombie and Batista-Navarro, 2018b) consists of speeches with two sets of polarity class labels: those derived from the speakers' division votes, as well a set of labels produced by human annotators. The size of that corpus was restricted to a little over a thousand examples due to the costs of manual annotation. In addition, only example speeches consisting of a maximum of five utterances were included in the corpus in order to make the task manageable for the annotators. The corpus has been used in sentiment polarity classification experiments by Abercrombie and Batista-Navarro (2018a) and Bhavan et al. (2019).

In other domains, the benchmark corpora that have been widely used in recent years run into the tens and even hundreds of thousands of examples (such as the *Yelp* dataset, which contains 500,000 reviews[1]). Until now, available resources in the domain of legislative debates have been significantly limited in their size by comparison.

4. Data

Transcripts of the parliamentary record, known as *Hansard*, are available in XML format at parliamentary monitoring website https://www.theyworkforyou.com/ under an Open Parliament licence.[2] This collection consists of all transcripts of debates from the House of Commons from 1919 to the present day, and is updated following each days' debates. We downloaded the most recent version of the transcript for each day from May 7th 1997 until November 5th 2019. This start date was chosen as it represents the beginning of the session of Parliament following that year's General Election. It is also the point at which speaker metadata began to be included in the record, enabling us to obtain the MP's names and party affiliations for inclusion in the corpus. The end date was the last day of the 2017-2019 Parliament.

We developed a tool to retrieve, for each debate, the motion(s) and the utterances of each speaker that voted in the corresponding division. We automatically omitted non-speech elements included in the transcripts such as '*[laughter]*' and '*rose—*', which are either presented in the transcripts between square brackets or are present in a list of such items that we had manually compiled.

We then automatically matched the debates to the corresponding divisions, which are presented in tables by TheyWorkForYou. For each speaker in each debate, and each motion, we matched the speaker's vote to their speech for use as a polarity label, with votes for '*aye*' and '*no*' representing *positive* and *negative* sentiment, respectively. In order to ensure that the vote labels correspond to the speeches in question, we retained only those debates for which we find exactly one motion and exactly one division. This left 34,010 example speeches in the corpus.

In addition, we obtained the speakers' names and party affiliations (at the time of the corresponding debates)

[1] https://www.yelp.com/dataset

[2] https://www.parliament.uk/site-information/copyright

from a resource maintained by TheyWorkForYou.[3] We matched these to the speaker identification numbers in the transcripts, and included the information as metadata in the corpus.

We present two versions of the corpus in CSV format: the full dataset (*ParlVote_full*) and the partially pre-processed subset that we use for our experiments (*ParlVote_concat*). In the former, each example consists of the following fields:

- *debate_id*
- *motion_speaker_id*
- *motion_speaker_name*
- *motion_speaker_party*
- *motion_text*
- *speaker_id*
- *speaker_name*
- *speaker_party*
- *label*
- *utterance_1*
- ...
- *utterance_n*

For the preprocessed dataset, we took the following steps to prepare the data. For each utterance, we removed all sentences containing the bigram '*give way*'. This procedural phrase features in many interjections in the House of Commons that consist of MPs requesting that the current speaker yield the floor, and was judged to indicate that the sentences in which it appears do not contain subjective language relating to the motion. In some cases, extraction of such a sentence led to the removal of the entire speech example in question. For each remaining example, we concatenated all the utterances into a single *speech*.

The original raw version of the corpus is composed of 34,010 example speech units, while the pre-processed version comprises 544 fewer speeches. With 52.91/47.09 per cent and 53.57/46.43 per cent *positive/negative* labels repectively, both versions have fairly balanced sentiment classes. See Table 2 for full corpus statistics.

The corpus is available for download at http://dx.doi.org/10.17632/czjfwgs9tm.1.

5. Experiments

5.1. Pre-processing

We tokenized the speeches and motions. Although lowercasing is a common pre-processing step for many NLP tasks, in this domain, casing may provide information about intended sentiment. For example, the words '*honourable*' and '*gentleman*' are likely to be positive (+0.71 and +0.125 mean sentiment scores, respectively, in sentiment lexicon SentiWordNet[4]), while in the House of Commons '*the Honourable Gentleman*' is an obligatory—and therefore neutral—procedural honorific'. We therefore omitted this step, keeping the texts' original casing.

	Full corpus	**Subset**
Speeches (example units)	34,010	33,461
Debates	1,995	1,995
Unique speakers	1,348	1,346
Parties	16	16
Max. parties per debate	11	11
Min. parties per debate	1	1
Mean parties per debate	3.63	3.61
Total tokens	25.74M	26.33M
Unique tokens	84.89k	81.59k
Max. utterances per speech	133	—
Min. utterances per speech	1	—
Mean utterances per speech	3.56	—
Max. tokens per speech	20,967	20,730
Min. tokens per speech	1	1
Mean tokens per speech	756.76	760.17
Max. tokens per utterance	7,431	—
Min. tokens per utterance	1	—
Mean tokens per utterance	212.61	—
Max. speeches per debate	154	149
Min. speeches per debate	1	1
Mean speeches per debate	17.05	16.77
Positive sentiment labels	17,993	17,721
Negative sentiment labels	16,017	15,740
Government motion examples	18,029	17,732
Opposition motion examples	15,981	15,729

Table 2: Statistics for the full *ParlVote* corpus and the pre-processed subset (in which all *utterances* have been concatenated into *speeches*, and some sentences containing procedural language have been removed) that we use for sentiment classification experiments.

5.2. Models

We used the dataset to evaluate the performance of a number of approaches to the automatic analysis of speaker sentiment in parliamentary debate speeches. We used combinations of the following text representations, machine learning classification methods, and approaches to modelling the debates:

Text representations

Bag-of-words We used unigram features as input to the classifiers, with term frequency-inverse document frequency feature selection.

Word embeddings We used pretrained BERT (Bidirectional Encoder Representations from Transformers) embeddings (Devlin et al., 2019), which we fine-tuned on the ParlVote data. With the intuition that casing carries important information in parliamentary debates, we used the *base, cased* model. We fine-tuned this model using the parliamentary data for three epochs (following the recommendations of Devlin et al. (2019)), and used the input to train

[3] Available at https://github.com/mysociety/parlparse/blob/master/members/people.json

[4] Baccianella et al. (2010)

neural network classification layers, as detailed below.

Machine learning classification methods

Support vector machine (SVM) Commonly used for sentiment analysis in this domain (Abercrombie and Batista-Navarro, 2018a; Balahur et al., 2009; Burfoot, 2008; Burford et al., 2015; Salah, 2014; Thomas et al., 2006; Yessenalina et al., 2010; Yogatama et al., 2015), this is a strong non-neural baseline. We used an SVM with a linear kernel, *L2* regularization, and a squared hinge loss function.

Multi-layer perceptron (MLP) A simple 'vanilla' neural network, which has been shown to perfom better than SVMs in some circumstances on this task (Abercrombie and Batista-Navarro, 2018a). We used a network with one hidden layer comprised of 100 nodes, batch normalisation, a ReLu activation function, a dropout regularization rate of 0.5, and sigmoid activation in the output layer. We used early stopping with a tolerance of three epochs to select the model used for classification of the examples in the test set.

Debate models:

Motion-independent: classification using features derived from the text of debate speeches only.

Motion-dependent:

- Two-step Government/Opposition motion-dependent classification. Abercrombie and Batista-Navarro (2018a) found that performance can be greatly enhanced by automatically separating those speeches made in response to Government-tabled motions from those directed at motions proposed by Opposition MPs, and classifying them separately. This is attributed to the fact that they tend to be *positive* and *negative* in sentiment respectively.

- Classification using text features derived from the target motions in addition to the debate speeches. This is an alternative approach to learning the effect of the contents of the motion on the speeches given in response to them.

In order to observe how well these systems perform when training on different quantities of data, and because the maximum sequence input size of BERT is 512 tokens, we tested each combination of classifier and debate model on the following five subsets of the corpus:

- *Large*

  The full pre-processed corpus subset, as described in Section 4.

- *Medium*
  - <= 512: All speeches/concatenated speeches + motions of 512 tokens or fewer.
  - *Any*: A random sample of examples of the same size as *medium (<= 512)* (18,253 examples).

- *Small*
  - *Any*: A random sample of examples of the same size as the corpus used by Abercrombie and Batista-Navarro (2018b) (1,251 examples).
  - <= 512: As above, restricted to speeches/speeches+motions of 512 tokens or fewer.

We evaluate these systems against a lower baseline of the majority class label in each training subset (the baselines therefore vary somewhat between subsets). All combinations were evaluated using the same randomly selected 80/10/10 training-validation-testing split of the data for each subsection of the corpus and each debate model.

6. Results

Results are presented in Table 3. We find that, with the exception of two classifier/debate model/data subset combinations, the machine learning classification methods outperform the majority class baselines. Both of these instances occur using the Government/Opposition motion-dependent debate model on the smallest subsets of the data, in which case there can be as few as 547 examples in a

Classifier	**Debate model**	**ParlVote corpus subset**				
		Small (1,251)		Medium (18,253)		Large (33,461)
		Any	<= 512	Any	<= 512	All
	Motion-independent	40.48	47.62	50.71	50.60	50.01
Majority class	Government/Opposition	53.17	53.17	49.62	52.00	49.09
	Motion + speech	40.48	47.62	50.71	51.86	50.01
	Motion-independent	50.00	52.38	59.26	55.48	61.78
SVM	Government/Opposition	51.59	**57.94**	**68.46**	**63.27**	66.24
	Motion + speech	50.00	59.52	60.51	57.78	61.82
	Motion-independent	50.00	53.97	59.69	54.87	60.05
MLP	Government/Opposition	46.83	55.56	66.54	63.00	65.34
	Motion + speech	44.44	56.35	60.84	57.67	62.18
	Motion-independent	48.41	50.79	57.39	54.38	60.56
BERT + MLP	Government/Opposition	**64.29**	53.17	66.70	61.25	65.61
	Motion + speech	57.94	50.00	61.39	60.30	**67.31**

Table 3: Sentiment classification accuracy scores (%) using five subsets of the ParlVote corpus. For each subset, the highest accuracy obtained is highlighted in bold text.

given training set, which may simply not be enough for the models to learn from.

The best performing combination of system and data is the SVM classifier on the medium *any*-length dataset using the Government/Opposition motion-dependent debate model, which obtains an accuracy of 68.46 per cent.

Classifiers While the machine learning approaches tend to beat the majority class baseline, in contrast to the findings of Abercrombie and Batista-Navarro (2018a), we find no consistent performance gains over the linear classifier from using a neural network.

Text representations Perhaps surprisingly—considering its success on other tasks and domains—we do not see consistent gains from fine-tuning on the BERT embeddings model, although this does obtain best performance on two of the corpus subsets.

Debate models Both motion-dependent models generally produce performance gains over the motion-independent speech-only model. These gains tend to be more prominent with the larger datasets.

Corpus size Classification performance generally improves as the amount of data increases, with all classifiers obtaining greater than 60 per cent accuracy on the large dataset. Limiting the length of the input does not appear to have the expected result of improving the performance of the BERT-based model.

7. Discussion

While it may be expected that a neural network, word embeddings-based approach such as BERT + MLP would outperform a somewhat simpler approach such as a linear SVM trained on a bag-of-words text representation, in most scenarios with this corpus, that does not appear to be the case. Given this, and the vastly quicker training time of the SVM (62.6 ms versus 1h 39min 48s for SVM/BERT + MLP on the *large* dataset in the motion-independent scenario running on a GPU), it may seem hardly worthwhile to pursue the transformer-based approach.

However, the most prominent setting in which the BERT + MLP model does produce the highest accuracy classification uses the largest dataset. In this scenario, the model appears to be able to take advantage of contextual information provided by the text of the motion, avoiding the need to train separately on Government- or Opposition-proposed motions (or indeed for the system to be provided with this information). As we only used a fairly simple, shallow model with standard parameter settings, there is certainly scope to experiment further with neural classification models for this task

In comparison to the short reviews and social media posts typically targeted for sentiment analysis, parliamentary debate speeches are inherently more complicated. While speakers must in theory address the proposed motions, the speeches can be long, cover diverse subject matters, include multiple targets of subjective language, and often feature irrelevant (to this task) procedural language. While we addressed the latter concern to some extent in removing sentences concerned with parliamentary turn-taking, manual observation of the transcripts reveals many examples of further off-topic and procedural language that remain in the corpus. Much room therefore remains for improvements that take account of these aspects of the debates when modelling them and selecting input for classification.

8. Conclusion

We have compiled and made available a new corpus of UK parliamentary debate speeches that is significantly larger than those previously available for speech-level sentiment analysis in this domain.

We tested the effects on classification performance of a range of combinations of machine learning methods, debate models, and corpus subset sizes. Results support prior work indicating that debate modelling should take account of motions to which speeches are addressed. They indicate that including text features from debate motions can lead to similar performance gains as the two-step models of Abercrombie and Batista-Navarro (2018a) over motion-independent classification.

Contrary to expectation, tailoring the examples to BERT's maximum input length did not lead to consistent performance gains. This may suggest that, for longer speeches, sentiment polarity can be determined from its first 512 tokens as well as it can from the whole speech.

The fact that all the accuracy scores are relatively low (under 70 per cent accuracy) indicates that this is a complex task, with plenty of scope for further analysis. Possible future directions include work on modelling the structure of the debates by, for example, further identifying and excluding non-sentiment carrying elements of speeches, such as procedural language.

9. Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments.

10. Bibliographical References

Abercrombie, G. and Batista-Navarro, R. (2018a). 'Aye' or 'no'? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Abercrombie, G. and Batista-Navarro, R. (2018b). A sentiment-labelled corpus of Hansard parliamentary debate speeches. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Balahur, A., Kozareva, Z., and Montoyo, A. (2009). Determining the polarity and source of opinions expressed in political debates. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 468–480, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bhatia, S. and P, D. (2018). Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 79–84, Brussels, Belgium, October. Association for Computational Linguistics.

Bhavan, A., Mishra, R., Sinha, P. P., Sawhney, R., and Shah, R. R. (2019). Investigating political herd mentality: A community sentiment based approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 281–287, Florence, Italy, July. Association for Computational Linguistics.

Burfoot, C. (2008). Using multiple sources of agreement information for sentiment classification of political transcripts. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 11–18, Hobart, Australia, December.

Burford, C., Bird, S., and Baldwin, T. (2015). Collective document classification with implicit inter-document semantic relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 106–116, Denver, Colorado, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ji, Y. and Smith, N. A. (2017). Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada, July. Association for Computational Linguistics.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Salah, Z., Coenen, F., and Grossi, D. (2013). Extracting debate graphs from parliamentary transcripts: A study directed at UK House of Commons debates. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pages 121–130, New York, NY, USA. ACM.

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

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, Sydney, Australia, July. Association for Computational Linguistics.

Yessenalina, A., Yue, Y., and Cardie, C. (2010). Multilevel structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1046–1056, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yogatama, D., Kong, L., and Smith, N. A. (2015). Bayesian optimization of text representations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2100–2105, Lisbon, Portugal, September. Association for Computational Linguistics.