Topic-centric sentiment analysis
of UK parliamentary debates

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

January 2021

By
Gavin Abercrombie
Department of Computer Science
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>13</td>
</tr>
<tr>
<td>Declaration</td>
<td>15</td>
</tr>
<tr>
<td>Copyright</td>
<td>16</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>17</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>18</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>20</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>21</td>
</tr>
<tr>
<td>1.2 Terminology</td>
<td>23</td>
</tr>
<tr>
<td>1.2.1 Sentiment analysis and opinion mining</td>
<td>23</td>
</tr>
<tr>
<td>1.2.2 Topics, policies, and policy preferences</td>
<td>24</td>
</tr>
<tr>
<td>1.3 Research questions, objectives, and contributions</td>
<td>25</td>
</tr>
<tr>
<td>1.4 Research contributions</td>
<td>27</td>
</tr>
<tr>
<td>1.5 Thesis structure</td>
<td>28</td>
</tr>
<tr>
<td>1.6 Published work</td>
<td>30</td>
</tr>
<tr>
<td><strong>I Background</strong></td>
<td>32</td>
</tr>
<tr>
<td>2 Parliamentary data</td>
<td>33</td>
</tr>
<tr>
<td>2.1 United Kingdom parliamentary debates</td>
<td>33</td>
</tr>
<tr>
<td>2.2 The Hansard record</td>
<td>37</td>
</tr>
<tr>
<td>2.3 Other data</td>
<td>38</td>
</tr>
<tr>
<td>2.3.1 Data from TheyWorkForYou</td>
<td>38</td>
</tr>
<tr>
<td>2.3.2 Data from other sources</td>
<td>40</td>
</tr>
</tbody>
</table>
5.1 Motion-dependent sentiment analysis with vote-derived and manual class labels ................................................................. 98
  5.1.1 Sentiment polarity class labels ......................................... 98
  5.1.2 Semantic Structure of House of Commons Debates .............. 99
  5.1.3 Corpus construction ..................................................... 101
  5.1.4 Debate speech sentiment models .................................... 107
  5.1.5 Experiments ............................................................... 109
  5.1.6 Results ................................................................. 110
  5.1.7 Discussion ............................................................... 112
  5.1.8 Conclusion ............................................................. 118

5.2 Sentiment classification using transformer-based language models 119
  5.2.1 Corpus construction ..................................................... 120
  5.2.2 Experiments ............................................................... 123
  5.2.3 Results ................................................................. 125
  5.2.4 Discussion and conclusion ............................................ 126

5.3 Chapter summary ............................................................ 127

6 Topic identification .... 129
  6.1 Topics in parliamentary debates ....................................... 130
    6.1.1 Pre-existing ontology of topics ................................ 130
    6.1.2 Topic modelling ..................................................... 130
  6.2 Multi-class, multi-label classification of motion policies ........... 132
    6.2.1 Opinion-topic labelling ........................................... 133
    6.2.2 Data ................................................................. 135
    6.2.3 Experiments ........................................................ 137
    6.2.4 Results ............................................................... 138
    6.2.5 Discussion ........................................................... 139
    6.2.6 Conclusion .......................................................... 140
  6.3 Policy preference detection in debate motions ......................... 141
    6.3.1 Background .......................................................... 141
    6.3.2 Data ................................................................. 142
    6.3.3 Annotation ........................................................... 144
    6.3.4 Experiments .......................................................... 147
    6.3.5 Results ............................................................... 149
    6.3.6 Discussion ........................................................... 152
  6.4 Chapter summary .......................................................... 153
7 Topic-centric stance analysis

7.1 Determining support or opposition for policy preferences

7.1.1 Data

7.1.2 Method

7.1.3 Results

7.1.4 Discussion

7.2 Chapter summary

III Conclusions

8 Conclusion

A HanDeSeT annotation guidelines

A.1 Introduction

A.2 Annotation procedure

A.3 Annotation guidelines for motion and speech sentiment polarity

A.3.1 Annotation examples

A.3.2 Explanation of label assignments

A.3.3 Dealing with ambiguity in motion stance

B Debate motion annotation guidelines

B.1 Introduction

B.2 Policies and coding

C Data statement for ParlVote+

D ParlVote+ annotation confusion matrix

E ParlVote+ annotation guidelines

E.1 Introduction

E.2 Policies and coding

E.3 Coding notes
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Extract from the Scottish National Party (SNP) manifesto 1997, as coded by Manifesto Project (MARPOR). Codes 301, 504, and 303 are Decentralisation: Positive, Welfare State Expansion, and Governmental and Administrative Efficiency, respectively.</td>
</tr>
<tr>
<td>2.2</td>
<td>Locations of the annotated corpora constructed for this project, and the sections of the thesis in which they are used.</td>
</tr>
<tr>
<td>3.1</td>
<td>Included publications by background and stated research objectives. Individual studies may have more than one objective.</td>
</tr>
<tr>
<td>3.2</td>
<td>Related work on topic identification in the legislative debates domain.</td>
</tr>
<tr>
<td>5.1</td>
<td>Inter-annotator agreement (Cohen’s kappa) for motion and speech sentiment polarity labels following the first and second versions of the annotation guidelines.</td>
</tr>
<tr>
<td>5.2</td>
<td>Causes of disagreement for the first round of annotation according to manual identification.</td>
</tr>
<tr>
<td>5.3</td>
<td>Occurrences of sentiment labels in the HanDeSet corpus.</td>
</tr>
<tr>
<td>5.4</td>
<td>Speech sentiment classification results: accuracy and F1 scores for combinations of debate model, motion and speech label types, and features (speech text, party, debate ID, and motion text) using SVM and MLP classifiers. Highest overall scores for each metric are highlighted in bold, best scores using text features only are underlined.</td>
</tr>
<tr>
<td>5.5</td>
<td>Top 10 most discriminating positive and negative n-gram features for motion sentiment classification, ranked by learned coefficients of the support vector machine (SVM) classifier using manually annotated labels.</td>
</tr>
</tbody>
</table>
5.6 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.

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

5.8 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 I use for sentiment classification experiments.

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

6.1 Output of structured topic modelling on the Hansard debates with \( n = 10 \) topics.

6.2 The number of motions labelled with each policy class label in the final dataset.

6.3 The numbers of policy preference codes, quasi-sentences (QSs) coded under each domain in the United Kingdom (UK) manifestos that I use as reference texts and training data, and debate motions and quasi-sentences that I label under each domain in the motion policy preference corpus.

6.4 The parties and years of publication of the manifestos that I use as reference texts and training data, and the number of labelled quasi-sentences (QSs) by party in this subset of the manifesto data.

6.5 Common interpretation of annotator agreement (Fleiss’s \( \kappa \)) at two levels of granularity during three iterations of training and development of annotation guidelines for labelling debate motions with the MARPOR codes.

6.6 Fleiss’ \( \kappa \) scores for three-way agreement and Cohen’s \( \kappa \) scores for two-way agreement on the debate motions dataset.
6.7 Statistics for the motion policy preference corpus. Gold standard examples have been labelled by two or three annotators initially and adjudicated on in a final round of annotation. Silver standard examples have been labelled by a single annotator only.

6.8 Domains of codes assigned to debate motions in the gold-labelled dataset.

6.9 F1 scores for classification of party political manifestos at the quasi-sentence level.

6.10 F1 scores for similarity matching and classification of debate motions at the quasi-sentence level. Word embeddings used were trained on Hansard (Dom. emb) or Wikipedia (Wiki emb).

7.1 Number of examples in the dataset labelled with each MARPOR policy preference code used. Codes used as class labels in the classification experiments described in Section 7.1.2 are highlighted in bold text.

7.2 F1 scores for classification of policy preference (multiclass), speech sentiment (binary), and policy-focused stance using motion-independent (Ind.) and motion-dependent (Dep.) debate models. Stance scores are reported as both the mean of the policy preference and sentiment scores and the absolute F1 score. The highest F1 scores for each task are highlighted in bold text.

7.3 F1 scores for classification of policy preference (multiclass), speech sentiment (binary) and policy stance (mean of these scores) using bag-of-words (BOW) and bidirectional encoder representations from transformers (BERT)-based text representations in the single-task-multi-layer perceptron (MLP) classification setting on speeches of 512 tokens or fewer.

7.4 F1 scores for policy preference, speech sentiment, and policy-focused stance detection by policy preference code label. The highest score for each task is in bold text. Contrastive pairs of policy preference codes are highlighted in grey boxes.

7.5 Mean sentiment scores for all speeches, supportive (+)/oppositional (-) speeches, replies to Government/opposition party motions, responses to own/other party motions, and all combinations of these three factors.
### 7.6 Top 20 discriminating features for the motion-independent setting

(all speeches), and, in the motion-dependent setting, responses to Government- and opposition-proposed motions, together with their mean SentiWordNet scores.

### 7.7 Number of speeches by token counts and prediction outcome

✓ = correct and ✗ = incorrect.

### D.1 Confusion matrix of policy preference annotations.
## List of Figures

1.1  Extracts from a House of Commons debate published in the Hansard record of October 22nd 2019, as presented on the parliamentary monitoring website TheyWorkForYou. Extract (a) is a debate motion, while (b) and (c) are utterances from supportive and opposing speeches, respectively. ........................................ 21

2.1  Screenshot from the Public Whip website showing one of its policies and some of the debates that have been categorized with that label by users of the site. ............................................................ 39

3.1  Flow diagram of the phases of the systematic review process: 1. database selection; 2. keyword search; 3. screening and eligibility assessment; 4. manual coding .................................................... 46

3.2  Coding framework for the included publications. ...................... 47

3.3  Rate of publication of papers featured in the review from 2003, the year of publication of the first paper, to 2018. The review additionally features one paper published in 2019. .................... 49

3.4  Global distribution of parliaments and legislatures from which debate transcripts are analysed in the studies included in this review. The size of each marker is proportional to the number of studies carried out on the legislature in question. ........................................ 51

3.5  Typology of sentiment and position-taking analysis tasks performed on legislative debate transcripts, showing the eight task types identified in this review. .................................................. 58

3.6  Machine learning methods used in the included publications for sentiment and position analysis. ....................................... 62

4.1  The supervised learning paradigm for text classification. ........ 79

4.2  Visual representation of a word embeddings vector space. ....... 82
4.3 A simplified representation of the process of fine-tuning a BERT model for classification. ........................................... 84
4.4 The maximum-margin hyperplane (thick, red line) of an SVM separating two classes. The example datapoints on the margins (dashed lines) are the support vectors. ........................................... 86
4.5 Architecture of a neural network with \( n \) input features \( \{x_1, ..., x_n\} \), two hidden layers \((l)\) of \( n \) nodes \( a_l^1 \) to \( a_l^n \), and an output layer that returns the predicted label \( \hat{y} \). ........................................... 88
4.6 Architecture of a convolutional layer in a convolutional neural network (CNN). ........................................... 90

5.1 Tree diagrams showing the number of example speeches by party affiliation, and how many of the associated class labels represent ‘rebel’ votes and speeches (inset) in which the sentiment labels support and oppose the majority of the speaker’s own party, respectively. ........................................... 107
5.2 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. ........................................... 108
5.3 Comparison of results using vote-derived and manually annotated class labels (using MLP classifier) in two conditions: with contextual metadata features excluded or included. ........................................... 113
5.4 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. ........................................... 114

6.1 Interpretation of policy labels for opinion analysis. In the PublicWhip data, for each of its policy labels, a motion has two further tags—outcome and policy vote—that can potentially reverse its opinion polarity. ........................................... 135
6.2 F1 score for different combinations of features used for topic classification. Highest performances (lighter, orange bars) are achieved when textual features derived from the debate titles are included. ........................................... 138
6.3 F1 scores for classification with different thresholds ($\theta$) for the minimum number of example motions. As this threshold is lowered, the number of topic classes ($n$) increases and performance decreases.

6.4 F1 macro scores for unigram overlap and cosine similarity matching at MARPOR’s policy and domain levels using textual features from whole motions. Use of cosine similarity leads to markedly better performance than unigram overlap, and the best performance is achieved using features derived from both the titles and motion texts at policy and domain levels.

7.1 Single and multi-task learning paradigms for classification of policy preference-focused speaker stance.

7.2 Distribution frequencies (histograms and density curves) of the correct and incorrect predicted probabilities of sentiment labels being positive for three categories of speech: those by all speakers, and intra- and inter-party responses.
Abstract

Parliamentary debate speeches provide access to the opinions and policy positions expressed by politicians towards many important topics. This information is of interest to citizens who wish to monitor the activities of their political representatives. However, due to the quantity, complexity, and specialised, esoteric language of the debates, they are not straightforward for human readers to process.

In prior work, sentiment analysis of legislative debates has been approached similarly to that applied to other domains. However, debate speeches are different in that their targets—the debate motions (proposals): (a) are non-neutral, which has a polarity-shifting effect on the content of speeches, and therefore on sentiment classifiers; and (b) are themselves sources of important topic information, without which, the output of analysis of the speeches is arguably uninformative.

I therefore examine the extraction of speaker sentiment with respect to the topics of the motions under debate. I evaluate state-of-the-art natural language processing (NLP) approaches to (1) sentiment polarity classification, (2) topic identification, and (3) topic-centric stance detection. These include the use of transformer-based language models, which I apply to this domain for (as far as I am aware) the first time. I compare approaches to class labelling for supervised classification, language representation, debate structure modelling, and machine learning methods and paradigms. The main contributions of this thesis are as follows:

**Sentiment polarity classification:** I evaluate approaches to this task, optimised for the domain of UK parliamentary debate speeches. I examine the validity of vote-derived sentiment class labels, finding that, to a large extent, they appear to align with the judgements of human readers. I propose a motion-dependent framework for dealing with the discourse structure of the debates, finding that this considerably boosts performance over motion-independent systems.

**Topic identification:** Topic-modelling yields overly broad outputs, which tend
not to be the targets of speech sentiment. Proposing instead a supervised approach, I evaluate labelling schema for this task. I explore the use of two labelling frameworks: debate motions labelled by crowdsourced annotators; and a schema designed by political scientists for the annotation of party-political documents.

Policy preference-focused stance detection: To answer the question ‘what is the position of speaker X on topic Y?’, I formulate this task as a form of topic-centric sentiment analysis. I evaluate a range of approaches to the task.

This work advances the state-of-the-art of sentiment analysis for the legislative debate domain. It provides insights into the nature of the task and remaining challenges, such as validation of commonly-applied assumptions about ground-truth speech sentiment, and analysis of sentiment-bearing parliamentary language.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/about/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

‘Thank you, fleshy beings of these united kingdoms! Exactly 125 of you will be spared when my Woopian Fleet arrives to lay waste to (what’s left of) your planet.’

Lord Buckethead, candidate for election as MP for Uxbridge & South Ruislip, 2019

Just as Lord Buckethead had a loyal core of supporters on whom he could depend to defeat rival candidate Count Binface by 125 votes to 69 (although ultimately losing out overall to Boris Johnson by 25,225 votes), I too have people to thank.

First and foremost, I am indebted to Riza Batista-Navarro, who has been an attentive and all-round great supervisor. I will miss working with you. Many thanks too to Goran Nenadic, who stepped in as substitute co-supervisor, and provided me with thoughtful and helpful advice. And thank you to my examination committee, Sara Tonelli and Andre Freitas, for taking the time to evaluate this thesis and provide an interesting and thought-provoking discussion during my viva.

Indispensible to this project have been the contributions of the annotators, without whose diligent work ploughing through hours of tedious debate transcripts, completion of this thesis would have been impossible: Kieran Lilwall, Loren Hosein, Melis Ince, Nancy Greig, Olga Sokolova, and Stefan Tasic.

Grazie mille to Federico Nanni, who was enthusiastic about my project, invited me to work in Germany, and has generally been extremely supportive. Thanks also to Simone Ponzetto, Goran Glavaš, and the team at the Data and Web Science Group at the University of Mannheim for making me feel so welcome as a visiting researcher.

And finally, I am grateful to my mother, Avril Powell, for putting up with me while I wrote this thesis during several months of pandemic-related lockdown.

Abbreviations

Adam  adaptive moment estimation
ANN  artificial neural network
BERT  bidirectional encoder representations from transformers
BOW  bag-of-words
CBOW  continuous bag-of-words
CNN  convolutional neural network
Con  Conservative Party
CS  computer science
CSV  comma-separated values
DUP  Democratic Unionist Party
ELMo  Embeddings from Language Models
EU  European Union
GPU  graphics processing unit
HTML  Hypertext Markup Language
ID  identifier
idf  inverse document frequency
Lab  Labour Party
LD  Liberal Democrats
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>latent Dirichlet allocation</td>
</tr>
<tr>
<td>L-BFGS</td>
<td>Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm</td>
</tr>
<tr>
<td>LSTM</td>
<td>long short-term memory</td>
</tr>
<tr>
<td>MARPOR</td>
<td>Manifesto Project</td>
</tr>
<tr>
<td>MLP</td>
<td>multi-layer perceptron</td>
</tr>
<tr>
<td>MP</td>
<td>Member of Parliament</td>
</tr>
<tr>
<td>NLP</td>
<td>natural language processing</td>
</tr>
<tr>
<td>NLTK</td>
<td>Natural Language Toolkit</td>
</tr>
<tr>
<td>OOV</td>
<td>out-of-vocabulary</td>
</tr>
<tr>
<td>PDF</td>
<td>Portable Document Format</td>
</tr>
<tr>
<td>QS</td>
<td>quasi-sentence</td>
</tr>
<tr>
<td>ReLU</td>
<td>rectified linear unit</td>
</tr>
<tr>
<td>SDLP</td>
<td>Social Democratic and Labour Party</td>
</tr>
<tr>
<td>SNP</td>
<td>Scottish National Party</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>tf</td>
<td>term frequency</td>
</tr>
<tr>
<td>tf-idf</td>
<td>term frequency-inverse document frequency</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>UKIP</td>
<td>United Kingdom Independence Party</td>
</tr>
<tr>
<td>ULMFiT</td>
<td>Universal Language Model Fine-Tuning method</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>UUP</td>
<td>Ulster Unionist Party</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

‘House of Commons language and procedure are virtually unintelligible.’

Paul Flynn, MP 1987–2019

Parliamentary debate transcripts provide citizens with access to important information, but are notoriously difficult for ordinary people to understand. To overcome this, automatic analysis of the opinions, positions, and policy preferences expressed in debate speeches can assist people to monitor the activities of their elected representatives, and thereby engage with the democratic process. In this thesis, I examine United Kingdom (UK) parliamentary debate transcripts, and the implications of their particular characteristics for the task of topic-focused sentiment analysis of House of Commons debate speeches. With this project, I seek to explore ways to extract information from the transcribed record that will allow for the analysis of the speech-giving activity of each Member of Parliament (MP) in relation to the topics discussed and policies proposed during debates. Specifically, I develop and evaluate methods for the following natural language processing (NLP) tasks in this domain: (1) sentiment polarity classification, and (2) opinion-topic identification, as well as (3) ways of combining the two to conduct topic-centric sentiment analysis.
1.1 Motivation

Politicians in the Parliament of the United Kingdom of Great Britain and Northern Ireland (the UK Parliament) directly represent single-Member constituencies and their constituents through their activities. Under this system, individual MPs are particularly accountable to their voters and the geographical areas they represent (Crewe 2015). One effect of this is that voters can consider themselves to have a direct line of communication with Parliament through their elected representative (Proksch and Slapin 2015). As a result, citizens may be interested in monitoring the parliamentary activities of their individual MPs to a greater extent than in other territories (where, for example, MPs may be chosen from a list at the larger regional or even national levels).

Figure 1.1: Extracts from a House of Commons debate published in the Hansard record of October 22nd 2019, as presented on the parliamentary monitoring website TheyWorkForYou. Extract (a) is a debate motion, while (b) and (c) are utterances from supportive and opposing speeches, respectively.
CHAPTER 1. INTRODUCTION

Written transcripts of the debates that take place in Parliament can provide access to the opinions and attitudes of MPs and their political parties towards the most important topics facing society and its citizens. The published transcripts, commonly known as the Hansard record, cover debates from 1802 to the present day, and are freely and publicly available in digitized formats. They are of interest to the politicians themselves, the media, scholars in fields such as political science and history, and any members of the public who wish to scrutinise the activities of their elected representatives. Figure 1.1 shows example extracts from a debate published in Hansard.

Despite this accessibility, it is difficult for people to process and make sense of these records. With over half a million speeches from some 23,000 debates, and new transcripts being produced in three separate debating chambers on a near daily basis, Hansard consists of an extremely large collection of complex text documents. Parliament is known for its archaic customs, conventions and procedures, and the language used is often impenetrable to the layperson (Onyimadu et al., 2013; Salah, 2014). There do exist manually curated aggregations of voting records in relation to different issues, such as those of parliamentary monitoring website TheyWorkForYou\footnote{1}. However, while UK MPs are known to be tightly constrained in their parliamentary voting behaviour due to the whip system\footnote{2}, they are generally freer to take a more independent approach in their speeches. Here, they may express more personal views or the views of their constituents, and are able to diverge to a greater degree from the majority party line (Benoit and Herzog, 2015; Proksch and Slapin, 2015). There currently exists no straightforward way of monitoring the speech-giving activity of MPs without the necessity for laborious manual examination of the debate transcripts, which even the politicians themselves can find ‘unintelligible’ (Flynn, 2012).

Sentiment analysis (also known as opinion mining\footnote{3}) is, broadly speaking, ‘the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes’ (Liu, 2012). This often takes the form of automatic identification of the polarity (usually positive or negative) of the position taken by the holder of an opinion towards such a target.

\footnote{1}{Available at \url{https://www.theyworkforyou.com/} accessed 7 July 2020.}
\footnote{2}{Whips are MPs with responsibility for maintainance of party voting discipline. See \url{https://www.parliament.uk/about/mps-and-lords/principal/whips/} accessed 7 July 2020.}
\footnote{3}{For a discussion of the terminology concerning sentiment analysis, see Section 1.2}
Sentiment analysis is one of the most active areas of research in the field of NLP, and a widespread need for political information has been offered as a motivation for the development of such technologies (Pang and Lee 2008). However, until recently, automatic analysis of the positions taken by speakers in legislative debates, and particularly those of the UK Parliament, has received relatively little attention from researchers. As I show in Chapter 3 that research which has been undertaken tends to view the speeches in Hansard as being similar to documents from other domains to which sentiment analysis has been successfully applied, such as product reviews. Researchers have therefore relied on the same or similar methods, ignoring both the unique discourse structure of the debates, and any potentially relevant and important information regarding the discussed topics, which are in fact the targets of the expressed opinions.

In this thesis, I examine the unique characteristics of text documents in this domain. I investigate how the content and structure of parliamentary debates differs from the types of text to which sentiment analysis is more commonly applied, and propose methods of optimizing approaches to the task for this domain. Unlike previous work, I also examine ways of adding important context to such automatic analysis by determining the targets of sentiment in the debate speeches: opinion-topics and policies. I explore ways of extracting information about the positions taken by MPs towards these targets. Successfully applied, such analysis could provide citizens with the means to more easily assess the positions taken, and the policies supported and opposed, by their parliamentary representatives, encouraging a greater degree of civic engagement among the population.

1.2 Terminology

In this section, I clarify the terminology used to describe the NLP tasks undertaken in this thesis.

1.2.1 Sentiment analysis and opinion mining

In the NLP literature, the terms opinion mining and sentiment analysis are often used interchangeably (for a discussion of this, see Pang and Lee (2008)). They are employed to describe both the specific task of determining a document’s sentiment polarity (that is, positive or negative, or sometimes neutral), as well as
the more general problem area of automatically identifying a range of emotional and attitudinal ‘private states’ (that is, non-observable, subjective states)\textsuperscript{4} such as ‘opinion, sentiment, evaluation, appraisal, attitude, and emotion’ (Liu 2012). In a survey, Yadollahi et al. (2017) list nine such different subtasks in the area of sentiment analysis.

For the purposes of this thesis, I use ‘sentiment analysis’ as a general umbrella term, encompassing any tasks concerned with the extraction of information relating to speakers’ opinions and expressed positions, and ‘sentiment polarity classification’ for the more specific, binary (or ternary) classification task.

Stance detection can be viewed as the task of determining the sentiment polarity of a piece of text towards a ‘given target of interest’ (Mohammad et al., 2016). I use this term to describe the work in Chapter 7 in which both the topic of a debate and the speakers’ attitudes towards it are identified.

A closely related research area is that of argumentation mining, in which analysis of the reasoning behind people’s opinions is conducted (Budzynska and Reed, 2019). For instance, in the examples in Figure 1.1, such a system may aim to output ‘get out of the European Union as speedily as possible’ and ‘something that we know will damage this country economically’ as arguments made by speakers (a) and (b) respectively. While this is an interesting direction for research on the Hansard transcripts, the scope of this thesis, following most prior work in this domain (e.g., Burfoot, 2008; Salah, 2014; Thomas et al., 2006), is limited to the task of identifying the opinions and stances of speakers (positive and negative stance towards the motion to progress the Bill in question).

1.2.2 Topics, policies, and policy preferences

In this thesis, I use the following terms in connection with the broad NLP task of topic detection: topic, policy, and policy preference. A topic is ‘the subject of a discourse, argument, or literary composition; a matter treated in speech or writing; a theme’ (OED, 2019b). As such, generic topics may be too general or neutral to be targets of sentiment expressed by the speakers who discuss them. In fact, initial research and experiments on topic detection suggested this to be the case, as shown in Chapter 6.

A dictionary definition of the term policy is ‘a principle or course of action adopted or proposed as desirable, advantageous, or expedient; esp. one formally

\textsuperscript{4}See Quirk et al. (1985).
advocated by a government, political party, etc.’ (OED, 2019a). However, on the parliamentary monitoring website the PublicWhip (5) this term is used to refer to sets ‘of votes that represent a view on a particular issue’, and it is in this sense that I use it in Section 6.2.

Policy preference is a term commonly used in political science to refer to the stated positions that politicians take towards different policy areas (Budge et al., 2001). The identification of these is the focus of Sections 6.3 and 7.1.

1.3 Research questions, objectives, and contributions

Research questions

I conduct this study with the overall aim of investigating the effectiveness of approaches to topic-centric sentiment analysis of UK parliamentary debates. To address the existing gaps in the research, I have formulated the following research questions:

RQ1: Related work

What approaches have been taken to the automatic analysis of speakers’ sentiment and position-taking in parliamentary and legislative debates? What challenges does this domain pose for the application of sentiment analysis?

RQ2: Sentiment polarity classification

How effective are general approaches to sentiment polarity classification when applied to the domain of UK parliamentary debates? What characteristics of the debates affect the performance of such systems? How can sentiment classification be optimized for this domain?

RQ3: Opinion-topic identification

What are the characteristics of the topics that are the targets of sentiment expressed in UK parliamentary debate speeches? How can these be labelled and automatically identified?

**RQ4: Topic-centric sentiment analysis**

What approaches and methods to topic-centric sentiment analysis are most effective for detecting the policy preferences of speakers in UK parliamentary debates? What are the remaining challenges associated with this task?

**Objectives**

In order to answer the above research questions, I have established the following research objectives:

**O1**: To establish the current state of approaches to sentiment analysis of parliamentary and legislative debates by conducting a systematic review.

**O2**: To assess the available corpora and data, and identify the particular characteristics of UK parliamentary debates.

**O3**: To design and construct annotated corpora and datasets for use in the evaluation of systems designed for sentiment polarity classification, multi-class topic classification, and topic-centric sentiment analysis of UK parliamentary debates.

**O4**: To develop and evaluate approaches and methods for (1) sentiment polarity classification, (2) topic identification, and (3) topic-centric sentiment analysis of UK parliamentary debate speeches that take account of and exploit the structure and characteristics of the documents in this domain.

**Hypotheses**

In Part II of the thesis, I present the results of empirical experiments designed to meet objective O4, as described above. There, I test the following research hypotheses:

**H1**: MPs’ votes are unreliable sentiment/stance class labels, as they do not always reflect the opinions and positions expressed in their speeches.
In Chapter 5, I compare vote-derived labels with those produced by human annotators, as well as the performance of classifiers trained with these two types of class label.

**H2** The polarity shifts caused by the discourse structure of parliamentary debates can be mitigated by applying a two-step, debate motion-dependent classification model.

In Chapter 5, I propose two approaches to this: two-step classification, and use of MP party affiliation metadata as a proxy for motion sentiment labels. I test this hypothesis both there and in Chapter 7.

**H3** Supervised machine learning classifiers can predict the manually applied opinion-topic labels of debates from textual (and metadata) features of the debate motions and speeches in a multiclass setting.

I test this hypothesis using two different labelling schemes in Chapters 6 and 7, evaluating the classifiers’ performance against naive baselines.

**H4** Classification performance on sentiment and stance detection of parliamentary debate speeches benefits from approaches to text representation and machine learning paradigm that have achieved state-of-the-art results in other domains, specifically contextual embeddings and artificial neural network (ANN) machine learning methods compared to SVM classifiers and BOW or non-contextual word embeddings.

In Part II, and in Chapter 7, I compare the classification performance of a range of neural network architectures and machine learning paradigms with SVM baselines. I also compare BERT contextual embeddings with static BOW and word embedding approaches to text representation.

### 1.4 Research contributions

With this thesis, I make the following research contributions:

- **C1**: Design and development of new annotated corpora for the evaluation of systems designed for sentiment polarity classification, opinion-topic identification, and topic-centric sentiment analysis of UK parliamentary debates.

- **C2**: A systematic review of literature concerning sentiment and position taking analysis of parliamentary and legislative debates.
C3: Development and evaluation of novel approaches for sentiment polarity classification optimized for the domain of UK parliamentary debate speeches.

C4: Development and evaluation of labelling schema and approaches to the identification of the policy preferences (opinion-topics) discussed in UK parliamentary debate motions (proposals).

C5: Development and evaluation of approaches to the novel task of policy preference-focused stance detection—a form of topic-centric sentiment analysis, in which speaker stance is determined with respect to proposed policies—in the domain of UK parliamentary debates.

1.5 Thesis structure

This thesis is composed of seven further chapters, which are organised into three parts. It is structured as follows:

Part I: Background

Chapter 2 describes the structure of UK parliamentary debates and the nature of the Hansard record of debate transcripts, as well as the other sources of data used for this project. Here, I also detail the annotated corpora that I have constructed for the evaluation of approaches to sentiment classification, topic identification, and topic-centric sentiment analysis.

Chapter 3 presents related work. This is comprised of a systematic review of literature concerning sentiment and position-taking analysis of parliamentary (and other legislative) debates, and a survey of other relevant prior work on similar tasks and in related domains.

Chapter 4 explains the methods used in the reported experiments. Here I describe the machine learning methods used in the experiments in Chapters 5, 6, and 7. I also describe the metrics used for evaluation of these methods, as well as statistical measures for the assessment of inter-annotator agreement that I used to assess the validity of the labelled corpora.
Part II: Experiments and analysis

Chapter 5 concerns empirical experiments on binary sentiment polarity classification of debates at the speech level. The experiments are presented in two parts. In the first of these, I compare the use of speakers’ votes and manual annotations as ground-truth labels for the construction of sentiment classifiers. Here, I also propose and test methods for handling the structure and characteristics of the debates. In the second part of the chapter, I test these methods on a large-scale corpus alongside state-of-the-art general approaches to sentiment classification that have been successfully applied in other domains.

Chapter 6 examines the nature of opinion-topics in parliamentary debate motions (tabled proposals), and presents the results of experiments on policy-focused topic identification using two different class labelling schemes for multiclass classification. In this chapter, I compare the use of different approaches to text representation in the parliamentary domain.

Chapter 7 builds on the results of the previous two experimental chapters to propose and test a system that aims to identify the policy preference-focused stance of debate speakers, a form of topic-centric sentiment analysis. For this, I propose and evaluate a multi-task learning approach, as well as a range of debate models, and feature selection and machine learning methods.

Part III: Conclusions

Chapter 8 concludes the thesis. Here, I review the contributions and findings of the research conducted in the previous chapters with reference to my research questions, objectives, and hypotheses. I also assess the remaining challenges in this area and suggest directions for future work.
1.6 Published work

Some of the work described in this thesis has been previously published in a number of peer-reviewed publications, as follows:

   The systematic literature review presented in Section 3.1 was originally published as this journal article.

   This workshop paper describes creation of the corpus used in Section 5.1.

   This conference paper presents work on sentiment polarity classification described in Section 5.1.

   This conference paper presents work on sentiment polarity classification described in Section 5.2.

This workshop paper presents the experiments on multi-label opinion-topic classification described in Section 6.2.


The work on policy preference detection described in Section 6.3 was originally presented in this conference paper.
Part I

Background
Chapter 2

Parliamentary data

‘What dreary pages of interminable talk ... not relieved by a single original thought, a single generous impulse, or a single happy expression! Why, Hansard, instead of being the Delphi of Downing-street, is but the Dunciad of politics.’

Benjamin Disraeli, Prime Minister 1868, 1874-80

In this chapter, I characterise the content and structure of UK parliamentary debates and describe the text domain of the Hansard record of debate transcriptions. I also detail the other sources of data used in this project, and describe the development of the novel annotated corpora that I created for this project.

2.1 United Kingdom parliamentary debates

UK parliamentary debate transcripts are considerably different in structure, content, and style to other domains on which sentiment analysis is more typically performed, such as product reviews and social media. In this section, I outline the characteristics of this domain, and the implications of these for sentiment analysis. I also explain the resulting choices made with regard to modelling of the debates.

Parliamentary debate is a highly formalized form of goal-oriented institutional discourse [Roberts 2013]. Although a number of other objectives may lay behind the speech-giving activities of speakers in Parliament, such as negotiation
CHAPTER 2. PARLIAMENTARY DATA

of MPs’s roles and status (Sarfo, 2016), their primary aim is to signal their position with regard to the tabled motion. Despite the name ‘debate’, participants rarely succeed in persuading (or even seek to persuade) other MPs of the superiority of their own positions. Rather, ‘parliamentary speech is primarily an act of position-taking’ (Proksch and Slapin, 2015), for which the intended audience is their constituents and voters (usually via the media).

In this work, I therefore view the pragmatic goal of all speech to be communication of the speakers’ positions towards the policies proposed in the debate motions. In order to fulfil this goal, MPs make use of deliberative rhetoric that focuses on the pros and cons of conflicting political solutions (Ilie, 2016). For the purposes of the binary sentiment and stance classification tasks investigated in this thesis, I therefore assume all language used by the speakers to be intended to support such polarised positions. As the tasks undertaken in this thesis are evaluated with respect to voting outcomes, this work focuses on analysis of debates at the whole-speech level, and does not take into account inter-sentence relations.

The UK Parliament consists of two Houses: the House of Commons and the House of Lords. The former is the superior legislative body, which houses elected constituency MPs, and is the target of most attention from the public, the media, and academics (Russell and Sciara, 2007). It is therefore the source of data and the focus of this study. Debates in the House of Commons ordinarily consist of the following elements:

**Motions** Each debate begins with a motion, a proposal tabled by an MP who then typically makes a speech supporting the motion. Motions start with the words ‘I beg to move, That ...’. They can be ‘substantive’, requiring the House to support or oppose a policy, piece of legislation, or state of affairs, as in Example 2.1:

That this House recognises the contribution that nationals from other countries in the EU have made to the UK; and calls on the Government to ensure that all nationals from other countries in the EU who have made the UK their home retain their current rights, including the rights to live and work in the UK, should the UK exit the EU.

Alternatively, they may be ‘general’, asking, in neutral language, MPs to
acknowledge that a particular topic has been considered by the House\textsuperscript{1} as in Example 2.2:

\emph{That this House has considered the current situation in Syria and the UK Government's approach.}  

(2.2)

For sentiment polarity classification in Chapter 5 and topic identification in Section 6.3, I included only the former motion type under the assumption that these are likely to be the targets of opinionated speech.

Within the substantive category are motions concerning the passing of legislation, such as Example 2.3:

\emph{That the Bill be now read a Second time.}  

(2.3)

As motions of this last type contain little information about the topic of a debate, I excluded them from the datasets that required manual annotation of motions (Sections 5.1 and 6.3). However, in the corpus used in Sections 5.2 and 7.1 I included all three types of motion in order to obtain a larger dataset, more representative of the Hansard record as a whole.

Debates may include any number of motions, which can relate to different elements of a proposed piece of legislation, such as Clauses or Amendments to a Bill. For example, counter-motions are often proposed in which the wording of the original motion is amended in a way that, if passed, radically alters or even reverses the position taken by the House (Rogers and Walters 2015). In this work, I perform analysis of only those debates that contain exactly one motion. This ensures that the analysed speeches have a known target.

\textbf{Speeches} Following movement of the motion, a number of MPs may speak—when invited or allowed to do so by the Speaker (the chief presiding officer of the House)—any number of times during a debate. The number and length of speeches in a given debate therefore varies depending on the decisions of the Speaker as well as the allotted time for the debate in question. Each speaking turn may be comprised of a short statement or question or a longer passage that is divided into paragraphs in the transcript.

\footnote{As described at \url{www.parliament.uk/about/how/business/debates}, accessed 8 July 2020.}
As in the widely used ConVote corpus of debates from the US Congress (Thomas et al., 2006), I 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.

Monitored by the Speaker, MPs are, at least in theory, prohibited from deviating from the subject of the motion in question (Rogers and Walters, 2015). In this thesis, I therefore work under the assumption that the target of all speech sentiment is that debate’s motion, and that each speech is either supportive (positive) or oppositional (negative) towards it. Examples 2.4 and 2.5 represent instances of utterances made in reply to motion 2.1. They are supportive and oppositional (as determined by the accompanying votes of the speakers), respectively.

\[\text{Does my hon and learned Friend agree that this situation requires leadership and a Prime Minister who will advocate in the best interests of every single individual in this country, EU national or otherwise? Will she share with me support for the First Minister’s statement on inclusivity and the need for leadership in this debate?}\] (2.4)

\[\text{Briefly, I completely agree with the first part of the hon and learned Lady’s motion, which I have read very carefully, in which she recognises the contribution made by EU nationals, but does she not accept that the first responsibility of the Minister for Immigration and the Prime Minister is to British citizens, more than 1 million of whom are in European Union countries? Their rights must be protected, but her motion is silent on their interests.}\] (2.5)

**Divisions** For each motion proposed, the Speaker asks the House to either agree or disagree to the motion. If the outcome of this decision can be ascertained simply by Members calling out whether they agree or not (near unanimous votes), no formal record of the vote is taken. Otherwise, a division (vote) is called. In such cases, the MPs then physically file through one of two division lobbies to have their votes recorded by four MPs (two on each side) who act as tellers.

Similarly to previous work (e.g., Salah, 2014; Thomas et al., 2006), I use the
record of these divisions as class labels for sentiment polarity classification of debate speeches—‘aye’ for the positive class, ‘no’ for negative. To ensure that these vote labels are attached to the relevant utterances, I include in the constructed corpora only those debates that feature exactly one division, discarding those debates in which multiple motions are voted upon. In Section 5.1, I investigate the link between these class labels and the ground-truth of speaker sentiment polarity, by comparing them with labels produced by manual annotation.

2.2 The Hansard record

A text domain with unique characteristics, the Hansard record lies somewhere between formal written language and transcripts of spoken dialogue. The transcripts are available in digital formats for parliamentary sessions from 1802 to the present day, and are updated on a rolling basis following each new day of debate.

The level of debate coverage in the record varies considerably over time. For modern editions of Hansard (since 1878), reporters have been employed to attend Parliament and record the debates.\textsuperscript{2} Parliamentary reporting has since seen further organisational and technical advances. Consequently, the level of integrity of Hansard has increased with time: while the record from the 1800s is sporadic, modern editions of Hansard represent almost complete accounts of events that take place in Parliament. In addition to these increases in the amount of available data, the style of reporting also changes diachronically. While early reports were written as third-party perspective summaries of debates, modern transcripts are largely-verbatim records of the speeches made in both chambers of the UK Parliament. That is, almost everything that is said in the chamber is transcribed, although repetitions and disfluencies are omitted, and some supplementary contextual information, such as speaker names, is added by the parliamentary reporters. In this project I focus on debates from 1997 onwards, as, from that year the record became most complete in terms of recorded speech, and the inclusion of metadata such as speaker names and identification numbers was initiated.

\textsuperscript{2}Such as, the assumption of responsibility for the production of Hansard by Parliament in 1909, and the various developments in the mechanical and electronic recording tools used by reporters.\textsuperscript{3}}
The Hansard transcripts are published under an Open Parliament Licence\(^3\) and are available online from two main sources: Hansard Online, the official record published by the UK Parliament, and TheyWorkForYou, a parliamentary monitoring website run by the charity mySociety\(^4\) which obtains the data from the former to publish on its site. Here, the transcripts are enhanced with information such as annotated notes and analysis of MP voting records. These two sources provide documents in different formats—Hypertext Markup Language (HTML) and Portable Document Format (PDF) at Hansard Online, Extensible Markup Language (XML) at TheyWorkForYou—and with varying degrees of coverage of the historical Hansard record. Hansard Online includes transcripts from July 1802 to the present day, but is incomplete, while the files at TheyWorkForYou comprise a complete record of debates from February 1919. Since some records are missing from Hansard Online (notably all transcripts from the year 2005 and much of 2006), for the construction of the corpora used in this project, I downloaded the data in XML format from TheyWorkForYou.

2.3 Other data

In addition to the debate transcripts, I have used several other sources of data to obtain both MP metadata information and labels for the training and evaluation of supervised sentiment, opinion-topic, and policy classification models.

2.3.1 Data from TheyWorkForYou

In addition to the speech contributions, the files obtained from TheyWorkForYou include the following data:

Speaker names, identification numbers, and party affiliations  Speakers’ names are included in the XML files under the attributes `speakername` or `personname`, where they may be referred to by their full name, title plus name (for example, ‘Mr. David Lammy’), position only (‘The Parliamentary Under-Secretary of State for Culture, Media and Sport’), or position plus title plus name (‘The Parliamentary Under-Secretary of State for Culture, Media and Sport (Mr.

\(^3\) Licence details available at [https://www.parliament.uk/site-information/copyright-parliament/open-parliament-licence/](https://www.parliament.uk/site-information/copyright-parliament/open-parliament-licence/), accessed 7 July 2020.

David Lammy). From 1997 onwards, speech contributions are provided with speaker identification attributes—named variously speakerid or personid. In order to resolve and standardise the identity of the speakers, and determine their party affiliations at the time of their contribution, I matched these attributes with entries in (a) separate lists of MPs from the TheyWorkForYou website and (b) another mySociety resource containing diachronic party membership information. I include this information in the datasets as metadata associated with each speech or motion.

**Division votes**  Divisions are included in the TheyWorkForYou transcript files as lists or tables (depending on the date) of the names of MPs who voted ‘aye’ and ‘no’. Because these appear in different formats, and as multiple MPs may share the same names, these require speaker identity resolution, as above. I used votes extracted from these divisions as sentiment polarity labels in the classification experiments in Chapters 5 and 7.

---

*Figure 2.1: Screenshot from the Public Whip website showing one of its policies and some of the debates that have been categorized with that label by users of the site.*

---

5Downloadable as comma-separated values (CSV) files from [https://www.theyworkforyou.com/mps/](https://www.theyworkforyou.com/mps/), accessed 7 July 2020.

2.3.2 Data from other sources

Crowdsourced opinion-topic labels For the topic identification experiments in Section 6.2, I used the crowdsourced ‘policy’ labels from parliamentary monitoring website the Public Whip. These take the form of lists of (the titles and dates of) debates that have been manually categorised by crowdsourced annotators as belonging to one of over 280 policies. These are defined as being ‘stated positions on a particular issue’. Figure 2.1 shows an example policy. To make use of this information, I matched the listed debates with those in the debate transcript files, applying the corresponding numbered policy code to the debate motion in the file as an opinion-topic label.

Party manifestos and manifesto codes For policy preference detection (Section 6.3 and Chapter 7), I use codes and annotated party political manifestos from the Manifesto Project (MARPOR). This consists of a coding scheme and a collection of party political election manifestos annotated by trained experts, with codes (class labels) representing such preferences. I use the coding scheme as a labelling framework for debate motions and the content of the manifestos themselves as references for an unsupervised approach to topic identification. Table 2.1 presents an example extract of a coded manifesto.

<table>
<thead>
<tr>
<th>Text</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>The provision of care in the community has a direct bearing on adequate local government funding, and statutory obligations imposed on local government must be backed up with financial provision.</td>
<td>301</td>
</tr>
<tr>
<td>So must any need to increase the involvement of voluntary organisations who are key facilitators in the provision of care.</td>
<td>504</td>
</tr>
<tr>
<td>The cost of democratic control of the Health Service through such Commissions will be considerably less than the present “internal market” which lacks both democracy and efficiency.</td>
<td>303</td>
</tr>
</tbody>
</table>

Table 2.1: Extract from the SNP manifesto 1997, as coded by MARPOR. Codes 301, 504, and 303 are Decentralisation: Positive, Welfare State Expansion, and Governmental and Administrative Efficiency, respectively.

---


8 There are 282 Public Whip policy categories at the time of writing in July 2020.

In Section 6.3 I use version 4 of the coding scheme. Organised under seven ‘domains’, the scheme comprises 57 policy preference codes, all but two of which (408: Economic goals and 000: No meaningful category applies) are ‘positional’, encoding a positive or negative position towards a policy issue (Mikhaylov et al., 2008). Indeed, many of these codes exist in polar opposite pairs, such as 504: Welfare State Expansion and 505: Welfare State Limitation. The manifestos that have been labelled are coded at the quasi-sentence level—that is, units of text that span a sentence or part of a sentence, and which have been judged by the annotators to contain ‘exactly one statement or “message”’ (Werner et al., 2011). I also use all fifteen of the annotated UK (including Northern Ireland) manifests from the Manifesto Corpus Version 2018-1 (Krause et al., 2018)—that is those that have been coded under version 4 of the coding scheme.

For Section 7.1 I use the updated version 5 of the guidelines, in which 12 of the codes have been divided into two or more subcategories.

### 2.4 Corpora

Using these data, I have constructed the labelled corpora presented in Table 2.2, which I make available for the research community, along with the tools I have developed in order to gather and process the data.

### 2.5 Chapter summary

In this chapter, I have described the structure of UK parliamentary debates and characterised the text domain of the Hansard record of debate transcriptions. I have explained my rationale for the choice of debate types analysed in this project, and the sources of class labels obtained for use in supervised machine learning settings. I have also provided descriptions and access to the novel annotated corpora that I created for this project.

---


11 While Northern Ireland returns MPs to the UK Parliament, distinct political parties operate there, and MARPOR treats it as a separate territory.


## Chapter 2. Parliamentary Data

<table>
<thead>
<tr>
<th>Corpus</th>
<th>URL</th>
<th>Purpose</th>
<th>Size</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>HanDeSet</td>
<td><a href="https://data.mendeley.com/datasets/xsvp45cbt4/2">https://data.mendeley.com/datasets/xsvp45cbt4/2</a></td>
<td>Sentiment polarity analysis</td>
<td>1,251 speeches</td>
<td>5.1</td>
</tr>
<tr>
<td>ParlVote</td>
<td><a href="http://dx.doi.org/10.17632/czjfwgs9tm">http://dx.doi.org/10.17632/czjfwgs9tm</a></td>
<td>Sentiment polarity analysis</td>
<td>33,461 speeches</td>
<td>5.2</td>
</tr>
<tr>
<td>Motion Policies Corpus</td>
<td><a href="https://data.mendeley.com/datasets/j83yzp7ynz/1">https://data.mendeley.com/datasets/j83yzp7ynz/1</a></td>
<td>Opinion-topic identification</td>
<td>592 motions</td>
<td>6.2</td>
</tr>
<tr>
<td>Motion Policy Preference Corpus</td>
<td><a href="https://madata.bib.uni-mannheim.de/308">https://madata.bib.uni-mannheim.de/308</a></td>
<td>Policy preference detection</td>
<td>473 motions</td>
<td>6.3</td>
</tr>
<tr>
<td>ParlVote+</td>
<td><a href="http://dx.doi.org/10.17632/czjfwgs9tm">http://dx.doi.org/10.17632/czjfwgs9tm</a></td>
<td>Policy preference support/opposition analysis</td>
<td>33,311 speeches</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Table 2.2: Locations of the annotated corpora constructed for this project, and the sections of the thesis in which they are used.
Chapter 3

Related work

‘There can be no doubt that the Parliamentary literature of this country is one of the most remarkable features of the intellectual development of the age in which we live.’

William Gladstone, Prime Minister 1868–74, 1880–85, 1886, 1892–94

In this chapter, I survey the available literature related to the NLP tasks of sentiment analysis and topic identification as applied to the domain of parliamentary and legislative debates. I present the results of a systematic review of sentiment and position-taking analysis in this domain (Section 3.1), I describe other relevant work in this area that was not retrieved by systematic search (Section 3.2), and I review related work concerning topic identification in this domain (Section 3.3).

3.1 Systematic review of sentiment and position-taking analysis of parliamentary debates

A version of the work presented in this section has previously been published as Abercrombie and Batista-Navarro (2020b).

In recent years, legislative debates, such as those held in the UK Parliament, have attracted the attention of researchers from diverse fields and research backgrounds. These include, on the one hand, computer scientists working in the field of NLP who have investigated the application and adaptation to the political sphere of methods developed for sentiment analysis of product reviews and blogs,
CHAPTER 3. RELATED WORK

and who have also tackled other related tasks in this domain, such as topic detection. In addition, political and social scientists, traditionally relying on expert coding for the analysis of such transcripts, have increasingly been exploring the idea of viewing ‘text as data’ (Grimmer and Stewart, 2013), and using computational methods to investigate the positions taken by debate participants.

As a result, a wide range of approaches to the problem of automatic debate transcript analysis have been adopted, with research on this problem varying widely in its aims and methods. Within this body of work, there exist many inconsistencies in the use of terminology, with studies in some cases referring to very similar tasks by different names, while in others the same term may mean quite different things. For example, while both Chen et al. (2017) and Kapočiutė-Dzikienė and Krupavičius (2014) attempt to classify debate speakers according to party affiliation, the former refer to this as ‘political ideology detection’, and the latter as ‘party group prediction’. Conversely, a single term like ‘sentiment analysis’ may be used to refer to, among other things, support/opposition detection (Thomas et al., 2006), a form of opinion-topic modeling (Nguyen et al., 2013), and psychological analysis (Honkela et al., 2014). The approaches adopted range from statistical analyses to predictive methods, including both supervised classification and unsupervised topic modeling. There are also contrasting approaches to modeling the textual data, the level of granularity of the analyses, and, for both supervised learning methods and the evaluation of other approaches, the acquisition and application of the labels used to represent the ground-truth speaker sentiment.

With regard to synthesis of the existing research on this topic, Kaal et al. (2014) assembled researchers from diverse fields to investigate the problem of text analysis in political texts, Glavaš et al. (2019) presented a tutorial addressing similar themes to this review, and both Hopkins and King (2010) and Monroe et al. (2008) discussed the general differences in the aims and objectives of social scientists and computer scientists when working on such problems. However, as far as I am aware, there has been no comprehensive written overview, systematic or otherwise, of research in this area to date. The aim, therefore, of this review is to bring together work from different research backgrounds, locating and appraising literature concerning computational sentiment and position-taking analysis that has been undertaken to date on the domain of parliamentary and legislative debate transcripts. I assess the research objectives, the types of task undertaken,
and the approaches taken to this problem by scholars in different fields, and present suggested directions for future work in this area.

3.1.1 Research questions

In carrying out this review, I aim to answer the following questions in order to ascertain the current state of research in this area:

- What are the research backgrounds of the authors of papers published in this area, and to what extent is the work multi-disciplinary?
- From which parliaments and other legislatures have debates been analysed?
- What are the objectives of researchers from different backgrounds working on sentiment and position analysis of parliamentary debates?
- What sentiment analysis task subtypes have been undertaken to conduct analysis of parliamentary debates?
- What approaches have been taken to sentiment/position analysis of parliamentary debates?
- What conclusions can be made about the reported performance and outcomes of the sentiment/position analysis systems that have been described?

3.1.2 Review scope and method

For this review, I followed the established systematic review guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Moher et al., 2009). The use of systematic review methodology to conduct this review has enabled the recovery of a substantial body of relevant work, but meant the exclusion of some potentially interesting studies. I was unable to include a number of known relevant results uncovered by an initial scoping search of the Google Scholar platform, which, due to the lack of transparency of its search algorithm, does not facilitate replication. While somewhat limiting in this sense, the decision to adhere to a systematic methodology provides a replicable and transparent method of synthesizing and summarizing the literature and identifying future research priorities. I discuss relevant publications not recovered by systematic search in Section 3.2.
I limited the search to publications concerning the automatic analysis of the sentiment, opinions, and positions expressed by participants in the transcripts of debates in parliaments and other legislatures, and also excluded any studies that do not report the results of empirical experiments. The review protocol pipeline is shown in Figure 3.1.

Figure 3.1: Flow diagram of the phases of the systematic review process: 1. database selection; 2. keyword search; 3. screening and eligibility assessment; 4. manual coding.
The review covers all literature retrieved by systematic search of five digital library databases and repositories (see Figure 3.1 (1)). These were selected as they provided high coverage of the results obtained by the prior scoping search. That search also provided the basis for the keyword search terms, which I developed to return results that included, as a minimum, all the relevant publications previously found (Figure 3.1 (2)).

All searches were conducted on January 31st, 2019. Following deduplication, screening and eligibility assessment, 61 studies have been included in the review. Using the NVivo qualitative data analysis software package [Richards 2005], I coded these according to (a) their research backgrounds, (b) the legislature and language of the debate transcripts analysed, (c) their stated research objectives, (d) the sentiment and position analysis tasks undertaken, (e) the approaches taken and methods used, and (f) the reported performances of the described sentiment/position analysis systems (see Figure 3.2).

---

1In order to reduce the number of results returned, I also added exclusion terms such as ‘big data’ and ‘twitter’, which I found did not prevent the retrieval of any known relevant publications.
3.1.3 Research backgrounds

I categorized the research background of each study according to the institutional affiliation(s) of its author(s) and the nature of its venue of publication, coding them as either computer science, political/social science, or multi-disciplinary. I consider a study to be multi-disciplinary if it (a) is written by authors from two or more research backgrounds, or (b) the paper is published at a venue associated with a different research background than that of its author(s)’ affiliations(s).

While, it is of course possible that the work classed as being from computer science (CS) or the social sciences, actually involves some level of inter-disciplinary collaboration that does not fit within this definition (for example, I did not investigate the authors’ academic histories), this is a straightforward yet systematic way of obtaining a general overview of the research community working in this area.

I found that over half the studies were written by researchers from a CS background ($n = 35$). Within this are researchers working on two kinds on problems. Firstly, there are those who approach the transcripts from a computational linguistics perspective, and whose work relates to properties of the language used such as argumentation structures and dialogue ([Duthie and Budzynska, 2018b], [Naderi and Hirst, 2016]). The second, larger group consists of work that can be characterized as belonging to the field of NLP [NLP] and whose work is more focused on tools and applications (e.g., [Ji and Smith, 2017], [Thomas et al., 2006]).

Political and social scientists authored less than half the number of included studies as computer scientists ($n = 14$), and just 12 studies involve multi-disciplinary research. Of these, seven involve both computer scientists and political or social scientists ([Kapočiutė-Dzikienė and Krupavičius, 2014], [Lapponi et al., 2018], [Rheault, 2016], [Rheault et al., 2016], [Rudkowsky et al., 2018], [Sakamoto and Takikawa, 2017], [van der Zwaan et al., 2016]), three collaboration between linguists and computer scientists ([Honkela et al., 2014], [Iyyer et al., 2014], [Nguyen et al., 2013]), and two that include researchers from three different fields ([Diermeier et al., 2012], [Nguyen et al., 2015]). According to the number of studies published on this subject annually, interest in this area has been increasing over time, particularly in recent years (see Figure 3.3).
3.1.4 Parliaments and legislatures

Nearly all the included studies focus on one single legislature for analysis, with only Sakamoto and Takikawa (2017) and Proksch et al. (2019) comparing their approaches (to the analysis of the level of polarization, that is, ideological division, in parliaments) on transcripts from two or more different chambers. The United States (US) Congress is by far the most popular legislature for analysis, attracting the attention of 31 of the studies. This can partly be attributed to the global power and influence of the US and of the English language, but is also explained by the widespread use by NLP researchers of the ConVote corpus (Thomas et al., 2006) as a benchmark dataset for the evaluation of sentiment analysis systems. Indeed, including its original authors, 17 of the publications use this dataset, 15 of which are written from a computer science background, with Hopkins and King (2010) (social science) and Iyyer et al. (2014) (multi-disciplinary) the exceptions. In some cases, ConVote is used alongside one or more other non-legislative datasets (such as product reviews) for the evaluation of text classification methods (Allison, 2008; Burford et al., 2015; Chen et al., 2017; Iyyer et al., 2014; Ji and Smith, 2017; Li et al., 2017; Martineau et al., 2009; Yogatama and Smith, 2006).
In fact, only a little over half (37) of the studies are exclusively concerned with the analysis of legislative debates. On the whole, political and social scientists seemingly prefer to construct their own datasets from the congressional record to suit their research aims, while Sakamoto and Takikawa (2017) (multi-disciplinary computational social science) also do so.

Following Congress, the next most analysed legislatures are the UK Parliament (n = 9) (Abercrombie and Batista-Navarro, 2018a,c; Duthie and Budzynska, 2018b; Onyimadu et al., 2013; Rheault et al., 2016; Salah, 2014; Salah et al., 2013a,b; Vilares and He, 2017). As none of the prior work on the UK includes publically released datasets for classification of sentiment, topics, or stance, I constructed new corpora for these tasks, as detailed in Chapter 2. The EU Parliament is the next most studied (n = 5) (Frid-Nielsen, 2018; Glavaš et al., 2017b; Honkela et al., 2014; Proksch and Slapin, 2010; Proksch et al., 2019). The French (Akhmedova et al., 2018; Lefait and Kechadi, 2010; Plantié et al., 2008) and Canadian parliaments (Ahmadalinezhad and Makrehchi, 2018; Naderi and Hirst, 2016; Rheault et al., 2019) all appear in three studies, while the California State Legislature (Budhwar et al., 2018; Kauffman et al., 2018) and the German Bundestag (Proksch et al., 2019; Rauh, 2018) are both analysed in two papers. The Austrian (Rudkowsky et al., 2018), Dutch (van der Zwaan et al., 2016), Lithuanian (Kapočiūtė-Dzikienė and Krupavičius, 2014), Norwegian (Lapponi et al., 2018), Czech, Finnish, and Spanish (Proksch et al., 2019) parliaments, Swiss Federal Assembly (Schwarz et al., 2017) Polish Sejm (Dzieciątko, 2019), Japanese Diet (Sakamoto and Takikawa, 2017), and the UN General Assembly (Baturo et al., 2017) are all utilised in only one study each (see Figure 3.4). It is notable that, thus far, research in this area appears to have been restricted to data from North America, Europe, and Japan.

Nearly all the included publications consist of analysis of parliamentary or legislative data in a single language. Exceptions include Sakamoto and Takikawa (2017) who use two corpora in different languages (English and Japanese), and Glavaš et al. (2017b) who use a multilingual dataset (German, French, English, Italian, and Spanish), as do Proksch and Slapin (2010) (English, French, and German translations). In the latter case, while the original data are multilingual (23 official languages of the European Parliament), the transcripts have been translated to these three languages. Similarly, for speeches not originally in English,
Baturo et al. [Baturo et al. (2017)] use official translations from the other official languages of the UN (Arabic, Chinese, French, Russian, and Spanish). By far the most prominent language is English, analysed in 55 studies. This is followed by French and German, used in six studies each, and a long tail of languages that only appear in one study each (Czech, Dutch, Finnish, Italian, Japanese, Lithuanian, Norwegian, Polish, Spanish, and Swedish).

### 3.1.5 Research objectives, tasks, and approaches

I investigated three aspects of the studies under review: the authors’ stated research objectives; the task types undertaken in order to achieve those aims; and the approaches taken to tackling those tasks. For the latter, I report the granularity at which analysis is undertaken, the methods used, and, where applicable, the labels used to represent ground-truth sentiment/position.

#### Objectives

I examined the principal stated objectives of each study in relation to the backgrounds of the researchers (see Table 3.1). These generally fall into two categories: (1) method evaluation, in which novel methods are presented and assessed, and the focus of study is the performance of the presented method or system; and (2) political analysis, in which an existing method is used as a tool in order to answer a political science research question, and the goal is interpretation of the
<table>
<thead>
<tr>
<th>Objective</th>
<th>Computer science</th>
<th>Political and social sciences</th>
<th>Multi-disciplinary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datasets/resources</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguistic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sokolova and Lapalme</td>
<td>2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dermech et al. (2012)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Frid-Nielsen (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jensen et al. (2012)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Owen (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Froisich and Slapin (2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rheault et al. (2016)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Schwarz et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Baturo et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kim et al. (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kaulfman et al. (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Owen (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Froisich and Slapin (2010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rheault et al. (2016)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Schwarz et al. (2017)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System/method performance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abercrombie and Batista-Navarro (2018a)</td>
<td></td>
<td>Honkela et al. (2014)</td>
<td></td>
</tr>
<tr>
<td>Abercrombie and Batista-Navarro (2018c)</td>
<td></td>
<td>Kapočiutė-Dzikienė and Nguyen et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>Akhmedova et al. (2018c)</td>
<td></td>
<td>Kruševčić (2014)</td>
<td></td>
</tr>
<tr>
<td>Balahur et al. (2009)</td>
<td></td>
<td>Froisich et al. (2019)</td>
<td></td>
</tr>
<tr>
<td>Bhatia and P (2018)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bonica (2016)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Budhirwar et al. (2018)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burfoot (2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burfoot et al. (2011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burlord et al. (2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen et al. (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duthie and Budzynska (2018)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dzieciatkó (2019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glaž et al. (2017b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ji and Smith (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lelait and Kechadi (2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. (2012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martineau et al. (2009)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naderi and Hirst (2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Onyimadu et al. (2013)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plante et al. (2008)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salah (2014)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salah et al. (2013a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salah et al. (2013b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thomas et al. (2006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vilares and He (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yessenalina et al. (2010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yogatama and Smith (2014a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yogatama and Smith (2014b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yogatama et al. (2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Included publications by background and stated research objectives. Individual studies may have more than one objective.
chosen system’s output. While the former is the focus of all CS publications and a few from political science (e.g., Bonica 2016; Frid-Nielsen 2018), the latter was found only in papers from the social/political sciences (e.g., Diermeier et al. 2012; Owen 2017). For some political science papers, even where the primary aim appears to be the former, a common approach is to combine these objectives, first presenting a text analysis method, and then illustrating its potential by employing it to answer research questions or test hypotheses, as in Hopkins and King (2010) and Proksch and Slapin (2010).

Although the work of computer scientists generally focuses on system evaluation, they often state secondary application objectives which encompass motivations relating to contributions to civic technology or the development of tools for social scientists. For example, Burfoot (2008) suggests that ‘tools could assist researchers in understanding the nuances of contentious issues on the web by highlighting areas in which different sites or pages agree and disagree’, while Budhwar et al. (2018) hope that their work will ‘give ordinary citizens a powerful tool to better organize and hold accountable legislators without the costs of physical presence and full-time lobbying representation’.

While the production of corpora and datasets are among the secondary contributions of many of the featured papers (e.g., Abercrombie and Batista-Navarro 2018c; Salah 2014; Thomas et al. 2006), in the cases of Lapponi et al. (2018) (linguistically annotated corpus) and Rauh (2018) (sentiment lexicon), this is their principal objective.

Hopkins and King (2010) claim that a fundamental difference between the objectives of computer scientists and political or social scientists is that, while the former are interested in making predictions about latent characteristics of individuals documents (such as sentiment polarity), the latter are more concerned with characterizing corpora or collections of documents as a whole, for example by the proportion of positive or negative examples it contains. This point is supported by Monroe et al. (2008), who agree that individual document classification is ‘inappropriate to the task’, because, they suggest, under this model, representation of the whole data generation process is backwards—where classification presumes that class labels are manifestations of the underlying latent phenomena of interest, the reality is the other way around: people first hold opinions or positions, and subsequently express them in speech or writing.

Despite this dichotomy, as can be seen in the next section, there do exist
cases of political scientists tackling classification (Proksch and Slapin, 2010) or computer scientists undertaking the scaling task from political science (Glavaš et al., 2017b).

Tasks

Within the overall area of sentiment analysis (as defined in Section 1.2), I found that the following eight types of tasks were performed in the studies:

**Agreement and alignment detection:** analysis of the similarity of the position taken by a speaker and another entity (another speaker, or a person or organisation outside of the debate in question), \( n = 5 \) (Ahmadalinezhad and Makrehchi, 2018; Duthie and Budzynska, 2018b; Kauffman et al., 2018; Naderi and Hirst, 2016; Salah et al., 2013a).

**Emotion analysis:** including emotion, anxiety and wellbeing analysis, \( n = 3 \) (Dzieciątko, 2019; Honkela et al., 2014; Rheault, 2016).

**Ideology and party affiliation detection:** in the literature, a speaker’s party affiliation is often used as a proxy for their ideological position. This may be performed as either topic modeling or classification, \( n = 14 \) (Balahur et al., 2009; Bhatia and P, 2018; Burfoot, 2008; Chen et al., 2017; Diermeier et al., 2012; Iyyer et al., 2014; Jensen et al., 2012; Kapočiūtė-Dzikienė and Krupavičius, 2014; Lapponi et al., 2018; Lefait and Kechadi, 2010; Li et al., 2017; Nguyen et al., 2013, 2015; Taddy, 2013).

**Opinion-topic analysis:** simultaneous extraction of topics and the speakers’ positions towards them, \( n = 4 \) (Abercrombie and Batista-Navarro, 2018c; Nguyen et al., 2013, 2015; van der Zwaan et al., 2016).

**Polarization analysis:** analysis aggregated at the legislature level of the extent to which debate is polarized and its speakers are ideologically divided, \( n = 2 \) (Jensen et al., 2012; Sakamoto and Takikawa, 2017).

**Position scaling:** positioning of speakers or parties on a scale of one or more dimensions (such as left–right), \( n = 11 \) (Baturo et al., 2017; Frid-Nielsen, 2018).
Sentiment/opinion polarity classification: binary or ternary analysis. As votes are frequently used as opinion polarity labels, this includes the task of vote prediction, $n = 28$ (Abercrombie and Batista-Navarro, 2018a; Akhmedova et al., 2018; Allison, 2008; Balahur et al., 2009; Bansal et al., 2008; Budhwar et al., 2018; Burfoot, 2008; Burfoot et al., 2011; Burford et al., 2015; Duthie and Budzynska, 2018b; Hopkins and King, 2010; Ji and Smith, 2017; Kauffman et al., 2018; Martineau et al., 2009; Onyimadu et al., 2013; Plantié et al., 2008; Proksch et al., 2019; Rauh, 2018; Rheault et al., 2016; Rudkowsky et al., 2018; Salah, 2014; Salah et al., 2013a,b; Sokolova and Lapalme, 2008; Thomas et al., 2006; Yessenalina et al., 2010; Yogatama and Smith, 2014a,b; Yogatama et al., 2015).

By far the most frequently undertaken task is sentiment or opinion polarity classification (although it is not always named as such in the literature). In the majority of cases this takes the form of learning from speech documents the predictive features of either speakers’ votes (e.g., Salah, 2014) or manually annotated ground-truth labels (e.g., Onyimadu et al., 2013). Polarity classification is particularly prevalent in the computer science studies (24 out of 29), but despite the previously discussed claims of Hopkins and King (2010) and Monroe et al. (2008) that the task is incompatible with the aims of social scientists, some political scientists and multi-disciplinary teams also tackle this task (Hopkins and King, 2010; Proksch et al., 2019; Rudkowsky et al., 2018).

As all the tasks undertaken concern the analysis of the positions taken by debate participants, there is considerable overlap between them. Furthermore, there is sometimes some discrepancy between the name given to a task and the actual task performed. For example, Onyimadu et al. (2013) refer to the task they perform as both ‘sentiment analysis’ and ‘stance detection’, although it could be said that they actually carry out only sentiment polarity classification as they do not specify a pre-chosen target, a requirement of stance detection (as defined by Mohammad et al. (2017)). Meanwhile, Thomas et al. (2006) refer to this task as ‘predicting support’, and Allison (2008), working on the same problem and the same dataset, calls it variously sentiment ‘detection’ and ‘classification’.
Although Rheault et al. (2016) consider their work to be a form of emotion detection, they actually perform a form of sentiment polarity analysis at the whole legislature level, while Akhmedova et al. (2018) simply refer to the task as an ‘opinion mining problem’. Other terms used to refer to this include ‘attitude detection’ (Salah et al., 2013a), ‘vote prediction’ (Budhwar et al., 2018), ‘emotional polarity’ measurement, ‘predicting the polarity of a piece of text’ (Yogatama and Smith, 2014a), ‘sentiment classification’ (Yessenalina et al., 2010; Yogatama et al., 2015), and simply ‘sentiment analysis’ (Proksch and Slapin, 2010; Rauh, 2018; Rudkowsky et al., 2018; Salah, 2014; Yogatama and Smith, 2014b).

In some cases, more than one task is investigated. For example, by switching party labels for vote labels, Burfoot (2008) use the same method to perform both sentiment polarity and party affiliation (or ideology) detection. Sentiment polarity analysis is often used as part of an NLP pipeline as a sub-task of a different opinion mining task, such as agreement detection (Salah et al., 2013a). Similarly, Kauffman et al. (2018) use sentiment analysis as a sub-task and the output scores as features for alignment detection, while Duthie and Budzynska (2018b) do similar for ethos detection, and Budhwar et al. (2018) aim to predict vote outcome using the results of sentiment polarity analysis as features for the task. Conversely, Burfoot (2008) applies the results of classification by party affiliation to predict speaker sentiment.

Balahuret al. (2009) combine polarity with party classification, a task that can be considered to be a form of ideology detection, but which they name ‘source classification’. Indeed, this is another task that suffers from a lack of clarity over terminology, with some studies considering party affiliation to be a proxy for ideology (Diermeier et al., 2012; Jensen et al., 2012; Kapočiūtė-Dzikiienė and Krupavičius, 2014; Taddy, 2013), while others do not make this connection, extracting information about speakers’ ideologies from their sentiment towards different topics (Bhatia and P, 2018; Chen et al., 2017; Nguyen et al., 2013), or training a model on examples that have been explicitly labelled by ideology, and not party membership (Iyyer et al., 2014). Yet others perform party classification, making no mention of the relationship between party membership and ideology (Balahur et al., 2009; Burfoot, 2008; Lapponi et al., 2018; Lefait and Kechadi, 2010). Alternatively, Abercrombie and Batista-Navarro (2018a) explicitly assume that membership of the same party does not guarantee homogeneity of ideologies, investigating intra-party differences of opinion and positions. Also concerned with
ideology, position scaling, which I code here as a separate task, can be performed on different dimensions, one of the most common being the left- to right-wing (ideological) scale.

The literature contains several efforts to simultaneously extract topics and speakers’ attitudes towards them (opinion-topic analysis). A common approach is to combine topic-modeling with some form of stance detection. Nguyen et al. (2015) used a supervised form of hierarchical latent Dirichlet allocation (LDA) to extract topics and polarity variables. Van der Zwaan et al. (2016) generated separate topic models for different grammatical categories of words in efforts to obtain this information. And Nguyen et al. (2013) performed supervised topic modeling to capture ideological perspectives on issues to produce coarse-grained speaker ideology analysis. Topic modeling was also undertaken by Sakamoto and Takikawa (2017), who used it to analyze polarization, a task also tackled by Jensen et al. (2012). Meanwhile, Vilares and He (2017) also performed opinion-topic modeling to extract speakers’ perspectives—‘the arguments behind the person’s position’—on different topics.

A number of other tasks which fit under the broader definition of sentiment analysis have also been tackled. Polarization analysis is undertaken in both Jensen et al. (2012) and Sakamoto and Takikawa (2017), who investigated changes in the extent to which language in Congress is polarized over time. Meanwhile, emotion detection is the subject of three studies. Dzieciątko (2019) classified speakers as expressing happiness, anger, sadness, fear, or disgust, while Rheault (2016) attempts to identify the level of anxiety exhibited by speakers, and Honkela et al. (2014) analysed a corpus of congressional speeches under the PERMA (Positive emotion, Engagement, Relationships, Meaning, and Achievement) model (Seligman, 2012). Agreement detection, an end in itself for Ahmadalinezhad and Makrehchi (2018) and Kauffman et al. (2018), has been used by Burfoot (2008), Burfoot et al. (2011) and Burford et al. (2015) to predict speaker sentiment, while Salah et al. (2013a) use agreement information to construct debate graphs. Finally, Naderi and Hirst (2016) automatically compared speeches with another type of labelled text (statements from online debates) to identify positive and negative framing of arguments.

Although there exist exceptions (see above), a notable difference between the focus of tasks undertaken by NLP researchers and social scientists is that the former tend to perform analysis with regard to the target of expressed sentiment.
(a specific proposal (e.g., [Allison, 2008]), piece of legislation (e.g., [Thomas et al., 2006]), topic (e.g., [van der Zwaan et al., 2016]), or other entity), while the latter generally analyse speakers’ aggregated speeches, ignoring the targets of individual contributions, and instead attempting to project actors onto a scale (such as left-right) ([Iliev et al., 2019; Kim et al., 2018; Laver et al., 2003; Lowe and Benoit, 2013; Proksch and Slapin, 2010; Schwarz et al., 2017]). Grimmer and Stewart (2013) note that this can be problematic as manual ‘validation is needed to confirm that the intended space has been identified’, and suggest automatic detection of relevant ‘ideological statements’ (or opinion-targets) as an important challenge.

For a full typology of tasks identified in this domain, see Figure 3.5. While there exists some work on opinion-topic identification, none of the identified publications tackle detection of the stances towards policies expressed in speeches, which is the ultimate aim of this thesis.

![Figure 3.5: Typology of sentiment and position-taking analysis tasks performed on legislative debate transcripts, showing the eight task types identified in this review.](image)

**Approaches**

I consider the granularity (level of analysis), methods, features, and ground truth labels used (in the cases of both supervised learning methods and evaluation of other methods) for each publication.

**Granularity** While, in other domains such as product reviews, sentiment analysis is typically carried out at the document, sentence, and aspect levels, here I
found a number of approaches to segmenting the transcripts for analysis, including breaking them down to the *sub-sentence* level and aggregating sentiment over entire corpora. There also exist differences in the terminology used to refer to these levels.

The vast majority of studies conduct analysis at the speech level ($n = 39$). However, ‘speech’ appears to mean different things in different publications, and in some it is not immediately clear just what the unit of analysis actually is.

A speech may be considered to be the concatenated utterances of each individual speaker in each debate ($n = 16$). Alternatively, analysis may be conducted at the *utterance* or ‘speech segment’ level (that is, an unbroken passage of speech by the same speaker) ($n = 24$), although Akhmedova et al. (2018) refer to these as ‘interventions’, and Bansal et al. (2008) as ‘individual conversational turns’. While several researchers who use Thomas et al. (2006)’s ConVote corpus claim to analyse ‘speeches’, the dataset (usually used unaltered) is in fact labelled at the utterance level. Similar use of terminology can be found in other work, such as Vilares and He (2017).

A further eight papers report analysis at the speaker level. That is, they consider a document to be the concatenation of all speeches given by the same representative (Bonica, 2016; Diermeier et al., 2012; Kauffman et al., 2018; Kim et al., 2018; Owen, 2017; Schwarz et al., 2017; Taddy, 2013).

Other approaches are to analyse speeches at the coarser political party (or bloc/coalition) level (Frid-Nielsen, 2018; Glavaš et al., 2017b; Proksch and Slapin, 2010; Proksch et al., 2019; Sakamoto and Takikawa, 2017; van der Zwaan et al., 2016), or the finer sentence (Duthie and Budzynska, 2018b; Naderi and Hirst, 2016; Onyimadu et al., 2013; Rauh, 2018) or phrase (Jensen et al., 2012) levels. Although Rudkowsky et al. (2018) detect sentiment in sentences, they aggregate these scores to provide speech-level results. Iyyer et al. (2014) break speeches down to both these levels, while Naderi and Hirst (2016) do so for sentences and paragraphs.

At the highest possible level of granularity, four studies consider sentiment over entire corpora. Dzieciatko (2019) and Rheault et al. (2016) aggregate sentiment scores for all speeches, presenting analysis of the Polish and UK parliaments respectively. Meanwhile, Honkela et al. (2014) compare the overall sentiment of

---

2In several cases, it has been necessary to contact the authors for clarification or to manually examine the datasets used to obtain this information.
the European Parliament transcripts with other corpora at the whole dataset level, and, in addition to party-level analysis, compare the polarity of Japanese and US datasets.

Finally, Ahmadalinezhad and Makrehchi (2018) consider each document to be a ‘conversation between two individuals’—that is, the two parties’ combined utterances—in order to classify these as being either in agreement or disagreement.

Overall, computer scientists tend to work at finer-grained levels (speech, speech segment, paragraph, sentence, or phrase), while in political science, the preferred units of analysis is the actor (or individual politician, whose contributions spanning a range of time are pooled together), which is the target of most work on position scaling, a task very much associated with that field. This confirms to some extent the assertion of Hopkins and King (2010) that, while computer scientists are ‘interested in finding the needle in the haystack, ... social scientists are more commonly interested in characterizing the haystack’. Exceptions, from the political and social sciences, are Iliev et al. (2019), and Hopkins and King (2010)—who actually propose a method of optimizing speech-level classification for social science goals, and from computer science, Glavaš et al. (2017b), who also tackle the position scaling problem.

**Methods**  A wide range of approaches are used, but these can be grouped into the following five main methods (of which some publications use more than one):

- Dictionary-based: using lexicons to assign sentiment scores, \( n = 16 \) (Ahmadalinezhad and Makrehchi 2018; Balahur et al. 2009; Budhwar et al. 2018; Chen et al. 2017; Duthie and Budzynska 2018b; Dzieciałko 2019; Honkela et al. 2014; Onyimadu et al. 2013; Owen 2017; Proksch et al. 2019; Rauh 2018; Rheault et al. 2016; Salah 2014; Salah et al. 2013a,b; Vilares and He 2017).

- Statistical machine learning: used to learn a predictive function based on the data, \( n = 45 \) (Abercrombie and Batista-Navarro 2018a,c; Ahmadalinezhad and Makrehchi 2018; Akhmedova et al. 2018; Allison 2008; Balahur et al. 2009; Bansal et al. 2008; Bhatia and P 2018; Bonica 2016; Budhwar et al. 2018; Burfoot 2008; Burfoot et al. 2011; Burford et al. 2015; Chen et al. 2017; Diermeier et al. 2012; Duthie and Budzynska 2018b; Dzieciałko 2019; Honkela et al. 2014; Onyimadu et al. 2013; Owen 2017; Proksch et al. 2019; Rauh 2018; Rheault et al. 2016; Salah 2014; Salah et al. 2013a,b; Vilares and He 2017).
CHAPTER 3. RELATED WORK


In general, the scaling task is approached using methods based on simple unigram counts (although a few make use of machine learning approaches (e.g., Diermeier et al. 2012, Hopkins and King 2010, Rheault 2016)). The scaling task comes in two varieties: supervised, in which target speeches (‘virgin texts’) are compared to reference texts, and unsupervised, such as the Wordfish package introduced by Proksch and Slapin (2010), and used by Schwarz et al. (2017), Glavaš et al. (2017b), in the only study conducted from an NLP perspective that takes on the position scaling problem so favoured by political scientists, use a combination of semantic similarity measurement and harmonic function label propagation, a semi-supervised graph-based machine learning algorithm.

In total, roughly three quarters of included studies (\( n = 45 \)) make some use of machine learning, and within this area there are a multitude of different
Machine learning methods (45)

<table>
<thead>
<tr>
<th>Supervised learning (40)</th>
<th>Graph-based (8)</th>
<th>Unsupervised learning (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification (39)</td>
<td>Supervised Topic Modeling (1)</td>
<td>Clustering (3)</td>
</tr>
<tr>
<td>Supervised Hierarchical LDA (1)</td>
<td>Minimum Cuts (7)</td>
<td>Neural Networks (1)</td>
</tr>
<tr>
<td>Naive Bayes (6)</td>
<td>Shared Nearest Neighbors (1)</td>
<td>Latent Dirichlet Allocation (7)</td>
</tr>
<tr>
<td>Logistic Regression (10)</td>
<td>Boosting (1)</td>
<td>Topic modeling (7)</td>
</tr>
<tr>
<td>Maximum Entropy (1)</td>
<td>Self-organising maps (1)</td>
<td></td>
</tr>
<tr>
<td>Naive Bayes Classifier (1)</td>
<td>Neural Networks (11)</td>
<td></td>
</tr>
<tr>
<td>Neural Networks (11)</td>
<td>Probabilistic methods (1)</td>
<td></td>
</tr>
<tr>
<td>Random Forest (2)</td>
<td>Support Vector Machine (29)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.6: Machine learning methods used in the included publications for sentiment and position analysis.

approaches (see Figure 3.6). These can be broadly categorised as supervised learning \( (n = 40) \), semi-supervised \( (n = 1) \), or unsupervised \( (n = 11) \) methods. Supervised methods are the preferred approach for text classification, and a wide variety of algorithms are used, including logistic regression \( (n = 10) \), naive Bayes \( (n = 6) \), decision trees \( (n = 3) \), nearest neighbor \( (n = 3) \), boosting \( (n = 1) \), a fuzzy rule-based classifier \( (n = 1) \), maximum entropy \( (n = 1) \), nearest class classification \( (n = 1) \). A further eleven studies make use of neural networks, which range in complexity from ‘vanilla’ feed-forward networks such as the multi-layer perceptron to convolutional and recurrent neural networks, including the use of long short-term memory (LSTM) units. Of these, six are concerned with sentiment polarity analysis and four with ideology detection, which as previously discussed, are highly similar classification tasks, performed with different class labels. The exception is Duthie and Budzynska (2018b), who used recurrent neural networks modules in their ethos mining task.

Rather than evaluating different classification algorithms, some work focuses on the use of different regularization methods (for logistic regression) (Yogatama and Smith, 2014a,b; Yogatama et al., 2015), while other approaches to improving performance of classifiers include Boosting (Budhwar et al., 2018), and Collective Classification (Burfoot et al., 2011).

Much of the work on unsupervised learning focuses on use of topic modeling methods, the majority of which are variations of the LDA algorithm (Blei et al.
The cross-perspective topic-model of van der Zwaan et al. (2016) generates two topic models from the data: one over nouns to derive topics, and the other over adjectives, verbs, and adverbs, which is intended to produce opinion information.

Additionally, seven studies model debates as networks of connected speakers and employ graph-based methods. Bansal et al. (2008), Burfoot (2008), Burfoot et al. (2011), Burford et al. (2015), and Thomas et al. (2006) all approach sentiment polarity classification as a minimum cuts graph partition problem. Chen et al. (2017) construct an ‘opinion-aware knowledge graph’, propagating known ideology information through the graph to infer opinions held by actor nodes towards entity nodes. For position scaling, Glavaš et al. (2017b) used similarity measurements as edges between documents, and propagated the scores of known ‘pivot texts’ to the other nodes.

Although ANNs have dominated NLP research in recent years, achieving state-of-the-art results in many tasks, in the legislative debates domain, their use remains relatively unexplored. In the experiments performed in Part II, I therefore test these against linear machine learning methods, as well as investigating the use of a multi-task learning paradigm for the task of speech stance detection.

**Language models and feature selection** Although analysis of debate transcripts necessarily utilises textual features derived from the speeches, there are a variety of approaches to how this is modelled, and which types of features are selected. In terms of language modeling, the majority of studies represent the text as bags-of-words. However, some add contextual information with use of word embeddings. While their use is generally restricted to studies from computer science, Rudkowsky et al. (2018) (multi-disciplinary) and Glavaš et al. (2017b) (computer science) explored their use for aims normally associated with the social sciences. None of the included publications make use of the contextual transformer-based language models (such as BERT (Devlin et al., 2019)) that have dominated the field of NLP in the last few years. I compare the use of BOW text representations with both static and contextual word embeddings in the experiments in Part II.

---

3That is, unstructured, unordered arrays of n-gram (term) counts.
CHAPTER 3. RELATED WORK

Relatively little linguistic or structural analysis is undertaken in these publications. An exception is Ji and Smith (2017), who found that the application of Rhetorical Structure Theory does not produce satisfying results in this domain because its ‘discourse structure diverges from that of news’, and Balahur et al. (2009), who extracted parse trees to determine the target of speech sentiment.

In addition to textual information, many studies also make use of metadata features of the speakers, as well as other features such as those derived from the structure of the debates. The following six categories of features are used:

- **Textual:** including word (all studies) or character (Kapočiutė-Dzikiénė and Krupavičius, 2014) n-grams, custom dictionary keyword features (Budhwar et al., 2018), words from particular grammatical categories (Iyyer et al., 2014), n-grams, custom dictionary keyword features (Budhwar et al., 2018), words from particular grammatical categories (Iyyer et al., 2014), word embeddings (Bhatia and P, 2018; Glavaš et al., 2017b; Iyyer et al., 2014; Ji and Smith, 2017; Li et al., 2017; Naderi and Hirst, 2016; Rheault, 2016; Rheault et al., 2016), sentence embeddings (Rudkowsky et al., 2018), and parse trees (Balahur et al., 2009; Iyyer et al., 2014; Ji and Smith, 2017).

- **Debate discourse features:** including citations (Burfoot, 2008; Burfoot et al., 2011; Lefait and Kechadi, 2010), interruptions and speech length (Budhwar et al., 2018), n-grams from other neighbouring sentences (Yessenalina et al., 2010), and utterance statistics (number and duration of speaker’s utterances) (Kauffman et al., 2018).

- **Speaker metadata features:** including bill authorship (Budhwar et al., 2018), debate and speaker IDs (Abercrombie and Batista-Navarro, 2018a; Salah, 2014), debate type (Lapponi et al., 2018), donations (Bonica, 2016; Kauffman et al., 2018), geographic provenance (Lapponi et al., 2018), party affiliation (Abercrombie and Batista-Navarro, 2018a; Burfoot, 2008; Kauffman et al., 2018; Rudkowsky et al., 2018; Sakamoto and Takikawa, 2017; Salah, 2014), and gender (Lapponi et al., 2018).

- **Polarity scores:** the output of opinion polarity analysis used as a feature for prediction of another phenomena (such as sentiment polarity (Bhatia and
CHAPTER 3. RELATED WORK

• Relational and knowledge graph features: features based on the relationships between different speeches or speakers, speakers and other entities (such as the targets of expressed opinion), or the measured similarity between speeches. (Burfoot, 2008; Burfoot et al., 2011; Burford et al., 2015; Chen et al., 2017; Thomas et al., 2006).

• Speaker vote: speaker’s votes for/against a motion or piece of legislation, used as a feature to identify their ideology (Kim et al., 2018; Schwarz et al., 2017).

A key decision for researchers is that of whether or not to include non-textual, metadata features, and the answer to this is usually driven by their research objectives. In some studies, particularly those from political science focused on position scaling, the object may be to examine intraparty differences or to compare speech and vote behaviour, in which cases features such as party affiliation or vote are the dependent variable under observation, and cannot be used as features for analysis. For classification, while some researchers compare performance with and without this additional information (e.g., Abercrombie and Batista-Navarro, 2018a; Salah et al., 2013a), others prefer to exclude them entirely in order to make their methods more generalizable to debates from other domains such as online debates, which do not have access to such information (Burfoot, 2008; Thomas et al., 2006). I investigate the use of metadata features in Chapters 5 and 6.

Ground-truth labels Depending on the nature of the task being tackled, for supervised classification methods, and in some cases, validation of unsupervised methods, several different data sources are used to represent the ground-truth sentiment or position. In most cases, researchers opt to make use of some pre-existing data, with the most common being the speakers’ roll-call or division votes (n = 21). Here, the most common approach is to consider each speaker’s vote on a given debate to represent the ground-truth position taken in their speeches, which may be analysed at the whole (concatenated) speech level (as in Salah (2014)) or broken down into smaller units (as in Thomas et al. (2006)),

4Terms used in the US Congress and UK Parliament, respectively.
whereby each vote label is attached to multiple examples. A general difference in approach is that, while CS studies use these as ground truth, political scientists tend to view speech and vote as unconnected, and even explicitly compare the two on this basis, as in Schwarz et al. (2017). Lowe and Benoit (2013) used human annotations for validation of the output of their scaling method. Whether or not votes are actually reliable as ground-truth is a matter of contention. Although some computer science studies assume this to be the case (e.g., Salah et al. 2013b; Thomas et al. 2006), Abercrombie and Batista-Navarro (2018a) who compared votes with human produced labels, concluded that it is the latter which more closely reflect sentiments expressed by speakers.

An alternative approach, is to use manually annotated labels. While some researchers have made use of already existing expert annotations, such as the Chapel Hill expert surveys\footnote{Chapel Hill surveys available at \url{https://www.chesdata.eu/}, accessed 8 July 2020.},\footnote{MARPOR website \url{https://manifesto-project.wzb.eu/}, accessed 8 July 2020.} (Glavaš et al. 2017b; Proksch and Slapin 2010), others produced labelled datasets specifically for their purposes. Onyimadu et al. (2013) and Rauh (2018) both had in-house annotators label speech sentences as being positive, negative, or neutral, while Rheault (2016) used crowd-sourced coders to label sentences as anxious or non-anxious. Rudkowsky et al. (2018) also used crowd-sourced labels, but for evaluation rather than training purposes, in their case to assess negativity detection. To create labels for the validation of their scaling of speakers’ positions towards a given topic, Frid-Nielsen (2018) had experts follow the MARPOR coding scheme\footnote{DW-nominate scores available at \url{https://legacy.voteview.com/dwnomin.htm}, accessed 8 July 2020.} to produce policy position labels, although the reliability of these is also controversial (Mikhaylov et al. 2008).

Other data used as ground-truth labels are the speakers’ DW-nominate scores (scores derived from congressional legislators’ voting records\footnote{Library of Congress website \url{https://www.loc.gov/crsinfo/}, accessed 8 July 2020.}),\footnote{Library of Congress website \url{https://www.loc.gov/crsinfo/}, accessed 8 July 2020.} (Diermeier et al. 2012; Nguyen et al. 2013), their constituency vote shares (Taddy 2013), ‘issue’ labels from the Library of Congress’s Congressional Research Service\footnote{Library of Congress website \url{https://www.loc.gov/crsinfo/}, accessed 8 July 2020.} (Bonica 2016), word perplexity (van der Zwaan et al. 2016), sentiment analysis scores obtained from prior experiments on the same data (Sokolova and Lapalme 2008), and party affiliations ($n = 13$). While the latter have been widely used as a proxy for speaker ideology (Bhatia and P 2018; Jensen et al. 2012; Li et al. 2017),\footnote{Library of Congress website \url{https://www.loc.gov/crsinfo/}, accessed 8 July 2020.} Hirst et al. (2010) suggest that party membership is actually a confounding factor...
for this task.

I investigate the validity of vote-derived sentiment class labels in Chapter 5.

Performance and outcomes

With the research reviewed here having such varied objectives and undertaking many different analysis tasks, it is not possible to directly compare the reported performances of the methods proposed. Nevertheless, in this section I attempt to summarize some conclusions of the included studies that are potentially relevant to future work in this area.

For classification, machine learning methods, and particularly neural networks, seem to outperform other approaches. Here, just as in other domains, such as product reviews, dictionary-based sentiment analysis methods appear to have been superceded by machine learning approaches. In a direct comparison, Salah (2014) found that machine learning classification methods outperform those utilising both generic and parliament-specific lexica, while Balahur et al. (2009) improved lexicon-based performance with the addition of a support vector machine classifier. Given this, and also considering the conclusion of Allison (2008) that ‘classifier choice plays at least as important a role as feature choice’, which learning algorithms should be selected for classification in this domain? In the work reviewed here, support vector machines, used in 29 of the studies, are the most popular option—both as a common baseline, and as a default algorithm choice. Although, in general, the last decade has seen an explosion in interest in deep learning methods, here we see relatively little use of neural network-based machine learning. Those studies that do directly compare the performance of such methods with other classifiers suggest a tendency towards better performance using neural networks (Abercrombie and Batista-Navarro, 2018; Budhwar et al., 2018; Iyyer et al., 2014; Li et al., 2017).

For position scaling, political and social scientists do not tend to place the same emphasis on performance metrics such as accuracy, preferring to make comparisons between output manual analyses in order to investigate theory-based hypotheses. Indeed, discussion of technical performance in these papers often focuses on whether or not computational text analysis is valid at all when compared with expert examination. In this respect, Diermeier et al. (2012), Frid-Nielsen (2018), and Laver et al. (2003) conclude that, with some caveats, it is a legitimate approach. Lowe and Benoit (2013) note that their method appears to position
some speakers on a different dimension to that of their expert analysis). In the one computer science paper to tackle this problem, Glavaš et al. (2017b) reported equally promising results on mono- and multilingual data, as well as superior performance using word embeddings over a bag of words model.

In the reviewed publications, a large range of feature types are extracted from the transcripts. Most studies rely primarily on the bag-of-words model, and for textual features, the benefits of adding higher-order $n$-grams (bi-, tri-grams, etc.) appear inconclusive. While Plantié et al. (2008) reported improved performance with the addition of bigrams to their feature set, Abercrombie and Batista-Navarro (2018a) did not see significant improvement with the use of bi- and tri-grams. With the most common method of $n$-gram feature selection being term frequency-inverse document frequency (tf-idf) weighting, Martineau et al. (2009), noting that inverse document frequency (idf) favours rare features, find that, for the relatively homogenous domain of a particular parliament’s transcripts, their alternative Delta tf-idf representation leads to better classification performance.

As we have seen, the appropriateness of using metadata features depends on the objectives of the research. However, if optimal classification performance is the goal and information regarding the speakers’ party affiliations is available, this has been found to be highly predictive of expressed sentiment (Abercrombie and Batista-Navarro, 2018a; Salah, 2014). Inter-document relationship information regarding agreement between speakers also assists in sentiment polarity classification, and has been applied successfully by Bansal et al. (2008), Burfoot (2008), and Thomas et al. (2006), as has network information (Burfoot et al., 2011; Burford et al. (2015)). The latter showed that it is possible to model these relationships for any dataset using $n$-gram overlap. In another approach to modeling debate structure, Balahur et al. (2009) used dependency parsing to find targets, which seemed to improve classification and help to balance results obtained in the positive and negative classes. While Iyyer et al. (2014) also reported success in using parse trees as features for classification with a recurring neural network, Ji and Smith (2017) did not find improvement in the parliamentary domain (although they did in news articles).

When it comes to representing ground-truth, votes are not necessarily indicative of the opinions expressed in speeches, but for speech-level polarity analysis they can be a convenient option. The results of computational analysis by
Schwarz et al. (2017) provided support for manual analysis in political science (Proksch and Slapin, 2015) to indicate that representatives position themselves differently in their speeches than in their voting behaviour. However, the relatively small difference between votes and manual annotations (less than four percent of their corpus) found by Abercrombie and Batista-Navarro (2018a), suggests that relatively small gains are to be had by investing in human labeling where other forms of class label are available.

A number of observations arise about the use of language in this domain. For the UK Parliament, Onyimadu et al. (2013) found that ‘compound opinions’, sarcasm, and comparative structures are all confounding elements for classifiers. In German, Rauh (2018) noted that ‘positive language is easier to detect than negative language’, while Salah et al. (2013a,b) made a similar observation for the UK Hansard transcripts. The latter study explains this phenomena as an artifact of the ‘polite parliamentary jargon’ used in Parliament. This point is also backed up by Abercrombie and Batista-Navarro (2018a), who observed that, the most indicative features, even of negative polarity, are words not typically thought of as conveying negativity. Where negative adjectives and verbs are present, Sokolova and Lapalme (2008) found that these are highly discriminative features.

Discussion and conclusion

Scope for further inter-disciplinary collaboration Considering the nature of the problem at hand—computational methods for the analysis of political text—it is somewhat surprising how little crossover can be found in this domain between ideas from CS and political science, and how seldom the methods used by researchers from these different fields are adopted by researchers from the other disciplines. As an explanation for this, Hopkins and King (2010), Monroe et al. (2008), Lowe and Benoit (2013) provided insights into the differing aims of the two fields. However, despite these differences, CS researchers may well be able to benefit from the theoretical expertise of political and social scientists, such as the rigourous labelling schema and expertly coded corpora already existing in the field. With this in mind, in Sections 6.3 and 7.1 I explore the use of a coding framework devised by political scientists for the labelling of debate topics.

Similarly, more political and social scientists could consider going beyond the simple bag-of-words n-gram language models they currently rely on to investigate the use of more advanced NLP methods of representing text and handling feature
sparsity in natural language, such as word embeddings. In Part \[ III \] I investigate the use of a range of these techniques, including the first (as far as I am aware) use of transformer-based embeddings in this domain.

**Standardization of terminology** A problematic issue that arises from surveying the work included in this review is the wildly inconsistent use of terminology, even within each of the research fields represented here. There is a clear need for greater agreement on which terms to use to refer to the affective targets of interest and the names of the tasks designed to analyse them, as well as the varying levels of granularity at which analysis is performed. These inconsistencies often mean that it is difficult—or even, without further investigation, impossible—for the reader to understand just what is done in a given study.

**More fine-grained level of analysis** Studies included in this review have approached analysis of legislative transcripts at a wide variety of granularities, from the phrase-level to comparisons aggregating sentiment over entire corpora. However, for the sake of convenience, and in order to make use of existing labels such as votes, the majority conducted analysis at the speech level, or even if they have done so at a more fine-grained sentence or phrase level, they tend not to consider the discourse structure of the debates. As Burfoot (2008) pointed out, parliamentary and legislative debates are complex, with many topics discussed and sentiment directed towards varying targets in ways that a document level classifier can struggle to identify. There is therefore room to develop more complex analyses, capable of recognising the relationships between entities and targets in fine-grained sections of the transcripts, perhaps using argument mining methods that harness theories from fields such as communication theory (e.g., Naderi and Hirst, 2016) or even philosophy (e.g., Duthie and Budzynska, 2018b) in order to explore the relationships between actors, opinion, targets and other entities in debates.

There have also been few attempts to link expressed opinion with topic information. While there have been some efforts to do so at the political party level (van der Zwaan et al., 2016), and as as a form of perspective analysis (Vilares and He, 2017), as well as by scaling on pre-defined topic dimensions (Owen, 2017), the majority of studies have simply conducted analysis of sentiment towards a target, such as a Bill or motion, the topic of which is unknown. In order to provide truly useful information, it may make sense to focus efforts on the extraction
of topic-centric opinions and to conduct analysis at the level at which different topics are found in the data. In order to address this, in Chapter 6 I evaluate methods of topic identification in debate motions, and in Chapter 7 I combine this with classification of speaker stance towards the identified topics.

Use of ground-truth labels While the majority of studies that focus on supervised classification have relied on votes as ground-truth labels, it is debatable whether these actually represent the target phenomena—the opinion or position taken by the speaker. Manual analysis in political science [Proksch and Slapin, 2015] certainly suggests that, in many legislatures, representatives express different positions in speech than in their votes, a point supported by Schwarz et al. (2017), who compared the scaling of speeches and votes. In Section 5.2 I therefore investigate the validity of vote-derived sentiment polarity class labels.

The computational analysis of sentiment and position-taking in parliamentary debate transcripts is an area of growing interest. While the researchers working on this problem have varied backgrounds and objectives, in this review I have identified some of the common challenges they face. With the majority of CS work focusing on unknown targets (Bills or debate motions, the topic of which is not assessed), and political scaling being conducted on very coarse grained scales (left-right, pro/anti-EU), there has thus far been little effort to direct efforts towards examining the targets of the opinions expressed. For the aims of both political scholarship and civic technology, what is required in many cases is identification of these targets, namely the policies and policy preferences that are discussed in the legislative chambers. In Part II I therefore explore such target-specific analysis.

3.2 Other work on sentiment analysis of parliamentary debates

In this section, I discuss literature concerning sentiment analysis of parliamentary debates that was not retrieved by systematic search in Section 3.1.2.

Several related publications were not included as they are only available via databases that cannot be searched systematically, such as Google Scholar and
Semantic Scholar. For speech-level sentiment analysis, [Rohit and Singh (2018)] created a corpus of labelled debates from the Indian Parliament and classified the examples as being ‘supportive’ or ‘against’ the issue in question, but without determining what this target actually is. In work preceding [Duthie and Budzynska (2018b)] (which is included and discussed in the systematic review in Section 3.1), [Duthie et al. (2016)] introduced the task of ethos mining in Hansard transcripts, before developing a schema of types of ethotic support and attack in [Duthie and Budzynska (2018a)].

In work similar to that of [Balahur et al. (2009); Iyyer et al. (2014); Kapočiūtė-Dzikienė and Krupavičius (2014)] (which has been discussed previously), Sapiro-Gheiler (2018) classified debate speeches by ideology and party. [Hirst et al. (2010)] conducted cross-parliament ideology detection experiments, finding that patterns of defence/attack and Government/Opposition confound the ability of classifiers to learn the target phenomena. I propose debate models to mitigate this in Chapters 5 and 7.

In political science, [Benoit and Herzog (2015)] used the supervised method of [Laver and Garry (2000)] to scale Irish debate speeches. This was an early case study and application of methods similar to much of the other work on this problem (e.g., [Laver et al. 2003; Lowe and Benoit 2013]), requiring manual selection of debates on a pre-defined topic.

Finally, one relevant study has been published since I conducted the systematic review. In the closest work to that presented in Chapter 5 of this thesis, [Bhavan et al. (2019)] developed the motion-dependent and motion-independent debate models proposed in Chapter 5.1, adding graph-based features to the textual features in order to conduct sentiment polarity classification on the HanDeSet corpus. Again, they did not consider the topics or policies that serve as the target of the speeches, as I do in Chapter 7.

### 3.3 Topic identification

Approaches taken to topic identification fall into two main categories: unsupervised methods such as topic modelling, and supervised classification.
3.3.1 Unsupervised topic modelling

In the former category, early examples of generative topic models include probabilistic latent semantic analysis (Hofmann, 1999) and latent semantic indexing (Papadimitriou et al., 2000). One of the most widely used methods is LDA (Blei et al., 2002), which is the basis of many variants of topic models. For example, the relational topic model (Chang and Blei, 2009) builds on LDA to model the links between documents as a binary random variable that is conditioned on their contents, while the structured topic model (Roberts et al., 2014) is an LDA variant that allows for the introduction of arbitrary document metadata to the input. In the initial experiments described in Section 6.1.2, I apply LDA topic modelling to Hansard transcripts, finding that the output of such models is too broad and general to be the target of sentiment in debate speeches, and is therefore unsuitable for the aims of this project.

3.3.2 Supervised classification

Efforts have been made to develop supervised forms of the topic modelling approaches described above (e.g., Mccauliffe and Blei, 2008), or to combine topic modelling with supervised topic classification (Blei et al., 2003). However, as noted above, the output of topic modelling approaches appears to be unsuitable for the aims of this project. Of more utility is supervised document classification, which is the approach I take to topic identification in the experiments in Sections 6.2, 6.3, and Chapter 7.

One of the most prevalent document classification tasks is in a fact a form of topic identification: text categorization of news articles. Here, classification is performed on documents from datasets such as the Reuters news text collection, which is labelled with topics such as ‘mergers and acquisitions’, ‘wheat’, and ‘earnings reports’. Early approaches to classifying such documents focused on knowledge engineering, which involves the manual crafting of classification rules. Such methods have the drawback of being both overly rigid in their categorizations and expensive, sometimes taking many person-years to develop, as in the case of the CONSTRUE system (Hayes and Weinstein, 1990) (which took 9.5 years of expert labour to build).

\footnote{Reuters news text collection available at \url{http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html}, accessed 8 July 2020.}
CHAPTER 3. RELATED WORK

Machine learning approaches to document classification include the application of Decision Trees and Naive Bayes (both Lewis and Ringuette, 1994). One of the most successful approaches has been SVMs, which have been applied to a range of text categorisation tasks, including news (e.g., Dumais, 1998; Joachims, 1998), question classification (e.g., Van-Tu and Anh-Cuong, 2016), and sentiment analysis (e.g., Pang et al., 2002). I use SVMs as a strong non-neural baseline in the experiments in chapters 5 and 6.

In recent years, neural network and word-embedding based approaches have been successful in achieving state-of-the-art performance for document classification on datasets such as the AG news corpus\textsuperscript{10} and in the domain of encyclopedia entries from the DBpedia ontology dataset of Zhang et al. (2015). These results have been obtained using methods such as Convolutional Neural Networks (CNNs) (Johnson and Zhang, 2015a,b; Zhang et al., 2015) and transformer-based fine-tuning of word embeddings (Devlin et al., 2019; Howard and Ruder, 2018), which I employ in the experiments in Part II.

3.3.3 Opinion-topic identification

In Chapters 6 and 7, I report the results of attempts to identify policies and policy preferences in parliamentary debates. These are forms of topic which act as the targets of the sentiment expressed in debate speeches. Opinion-topic detection is a sub-task of subjectivity detection and has been the focus of considerable attention.

Early approaches to this task focused on lexical approaches, whereby systems simply looked up words related to products and their attributes in domain-specific lists (Hu and Liu, 2004; Kobayashi et al., 2004; Popescu and Etzioni, 2005; Yi et al., 2003). These methods were devised for the analysis of product reviews, and were not easily generalisable to other domains.

A more general approach was proposed by Kim and Hovy (2006), who used semantic frames from FrameNet\textsuperscript{11} to map subjective expressions to their targets in the news domain. Noting that the target spans identified using this approach do not always coincide with actual opinion-topics, Stoyanov and Cardie (2008) redefined these targets as consisting of target spans—which comprise the contents


\textsuperscript{11}FrameNet available at http://framenet.icsi.berkeley.edu/, accessed 8 July 2020.
of an opinion—and potentially multiple topic spans—entities—within these, and tackled the classification task as a topic coreference resolution problem.

A large body of work also exists on identifying the aspects or ratable attributes of target entities (such as the food, decor and service of a restaurant) for fine-grained analysis of the sentiment directed towards them (see Schouten and Frasincar (2016) for a survey).

While the methods discussed here could prove useful for a more fine-grained analysis, the aim of this project is to identify only the overall opinion-topic of speeches, which in UK parliamentary debates is defined by the motion under discussion.

### 3.3.4 Topic identification in parliamentary debates

Although there exists work on detecting topics in related domains such as political campaign speeches (e.g., Menini and Tonelli 2016; Nanni et al. 2016; Sim et al. 2013) and electoral manifestos (e.g., Glavaš et al. 2017; Menini et al. 2017; Zirn 2016; Zirn et al. 2016), the identification of topics in parliamentary debates has received relatively little attention. Table 3.2 presents a summary of the published work in this area.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Hansen et al. (2019); Lehmann (2019); Zirn (2014)</td>
</tr>
<tr>
<td>Clustering</td>
<td>Nanni et al. (2016)</td>
</tr>
<tr>
<td>Semantic tagging</td>
<td>Alexander and Davies (2015); Coole et al. (2020)</td>
</tr>
<tr>
<td>Topic modelling</td>
<td>Watanabe and Zhou (2020)</td>
</tr>
</tbody>
</table>

Table 3.2: Related work on topic identification in the legislative debates domain.

Some work has been undertaken on providing parliamentary corpora with additional semantic information. Alexander and Davies (2015) and Coole et al. (2020) both produced tagged Hansard corpora in which semantic information is added at the individual word level, but not providing document-level topic analysis. Nanni et al. (2019) produced a topic-annotated dataset of parliamentary debates. Like my work in Part III they considered the debate motions to be key to determining the subject of debates, and extracted key words from these which they considered to represent ‘key-concepts’ of the debates. Unsupervised clustering of these then produced topics such as ‘finance’ and ‘Northern Ireland’, but these required manual interpretation, and, while potentially useful for many
purposes, do not represent the type of opinion-topics that are targets of parliamentary speech sentiment.

For topic classification, Hansen et al. (2019) constructed a corpus of speeches from debates in the Danish parliament, and, using Naive Bayes and SVMs to learn classifiers, achieved impressive results in automatically labelling unseen examples of these. The class labels they applied are expert devised subject areas such as ‘Agriculture’, ‘Business’, and ‘Culture’. Similarly, Zirn (2014) classified debates from the German Bundestag as belonging to one of a fixed set of topics derived from the names of parliamentary committees such as ‘Affairs of the European Union’, ‘Labour and social Affairs’, and ‘Food, Agriculture and Consumer Protection’, and also proposed a topic modelling approach, outputting topic terms that require human interpretation. Also focusing on the Bundestag, Lehmann (2019) used supervised learning methods to classify both manifesto extracts and debate speeches of parties and politicians into similarly derived broad ‘policy domains’. Although they used the policy preference coding scheme from the Manifesto Project (see Section 1.2.2) for validation purposes, it was not used to directly label debates for training and testing purposes, as I do in Section 6.3. They also compared the results of scaling of (pooled) manifesto extracts and debate speeches on dimensions derived from the policy preference codes, but analysis was performed at the level of parties rather than individual speakers.

And Watanabe and Zhou (2020) used a seeded version of topic modelling to classify speeches from the United Nations General Assembly as belonging to one of the six topics ‘greeting’, ‘UN’, ‘security’, ‘rights’, ‘democracy’, and ‘development’. Although these approaches are somewhat similar to that proposed in Chapters 6 and 7, in both cases the classes are generic political subject areas, rather than the opinion-topic sentiment targets sought in this project.

3.4 Chapter summary

In this chapter, I have surveyed the available literature related to the natural language processing tasks of sentiment analysis and topic identification of parliamentary and legislative debates. I have presented the results of a systematic review of sentiment and position-taking analysis in this domain, described relevant work in this area not retrieved by the systematic search, and reviewed related work concerning topic identification in this domain.
My findings from this chapter include the following:

- Speakers’ votes are often used as class labels for supervised classification, although it is not clear to what extent they represent speech sentiment.

- Although researchers from both political and computer science are active in computational analysis of legislative debates, there appears to have been relatively little cross-polination of ideas from these fields.

- There has been relatively little use of the NLP methods that have achieved state-of-the-art results in similar tasks in other domains, such as ANNs and word embeddings.

- The approaches taken to sentiment analysis of legislative debates have not taken into account the targets of the sentiment present in the speeches.

- There are no prior publically available labelled datasets available for the tasks undertaken in this thesis.

These findings have informed the direction of the experiments and analyses conducted in Part II, in which I explore the validity of vote-derived sentiment class labels (Section 5.1), the incorporation of political science theory (Section 6.3), a focus on topic-centric sentiment (Chapters 6 and 7), and the application of state-of-the-art machine learning and text representation methods (Chapters 5, 6, and 7). As detailed in Chapter 2, I have created labelled corpora for these tasks, which I make available for the research community.
Chapter 4

Methods

‘Are we doomed to a cold and heartless future in which computer says ‘yes’ or computer says ‘no’ with the grim finality of an emperor in the arena?’

Boris Johnson, Prime Minister 2019–

The experiments presented in Part II concern the evaluation of NLP and machine learning methods for the analysis of text documents. In this chapter, I motivate the use of the chosen methods for the tasks of sentiment, topic, and stance detection in the domain of parliamentary debates. I describe these methods, as well as various standard approaches to text representation and feature selection, machine learning methods, performance metrics, and statistical annotator agreement measures, which I also employ in Part II.

Document classification The majority of the experiments in Part II concern various forms of automatic document classification. For text documents, the input to a classifier is a vectorized numeric representation of the document collection (Section 4.1 concerns approaches to creating such representations). In a supervised classification scenario, the documents are labelled with the target classes (for example, positive or negative sentiment labels). This is done either by human annotators or with automatically retrieved labels that are assumed to represent the phenomena of interest, such as use of the votes of the debate participants as a proxy for speaker sentiment. In the work described in this thesis, these class labels are binary in the case of sentiment classification (for example,
1 for \textit{positive} and 0 for \textit{negative}), and multi-class in the opinion-topic/policy preference classification tasks.

![Figure 4.1: The supervised learning paradigm for text classification.](image)

Figure 4.1 shows the standard supervised text classification paradigm. For evaluation of a classification system, the document collection is divided into training and test (and possibly validation) sets. The ratio of these depends on the size of the dataset in question and the distribution of the class labels within it. Where the size of the document collection is relatively limited (as in Section 5.1), in order to maximise the available data for training and testing of the models, I employ \( k \)-fold cross validation. In this scenario, the dataset is split into \( k \) subsamples, and the classifier is iteratively trained on \( k - 1 \) of these while using the remaining subsample as the test set.

To train a classifier, features are extracted from the documents in the form of numeric representations of the documents. Such representations from the training set are input to a learning algorithm, which iterates over the examples seeking a \textit{decision boundary} which optimally separates the datapoints in the different classes, while minimizing the error in classifying them. Multiple passes are made over the data until convergence—that is, until no further improvements in the error rate occur. The resulting parameters of the decision boundary constitute a classifier which can then be used to label the unseen test examples. The output labels of this test set can then be compared to the ground-truth labels in order to evaluate the performance of the model (see Section 4.4). In the following sections of this chapter, I describe the text representation and machine learning methods that I have used in greater detail. I also outline the performance metrics used to evaluate them, as well as statistical measures of inter-annotator agreement.
**Topic modelling** In addition to text classification methods, in Section 6.1.2 I consider the use of generative statistical models to identify topics in parliamentary debates. In Section 4.3 I describe the topic modelling method I use.

### 4.1 Text representations and feature selection

In order to process natural language as data, it is necessary to represent the texts numerically. I use the following two approaches to this step, each of which have their own characteristics, advantages, and disadvantages.

#### 4.1.1 Bag-of-words with term frequency-inverse document frequency weighting

A *bag-of-words* (BOW) model is an unordered set of the words present in a collection of text documents. For each word $w$ in the vocabulary of a corpus or collection $D$, a binary feature indicates the presence (1) or absence (0) of $w$ in each document $d$ (Jurafsky and Martin [2014]).

This model has the disadvantage of discarding the structure and word order of a document (although direct collocations can be preserved with the inclusion of multi-word items (*bigrams, trigrams*, etc.) in addition to single words). Another drawback of this approach is that a BOW also erroneously represents all vocabulary items as being independent of one another. In reality, the likelihood of a particular word appearing in a particular document is likely to depend heavily on the subject-matter of the text, and hence the words with which it appears. Nevertheless, BOWs have been shown to work well on many NLP tasks (Manning et al. [1999]).

When representing a dataset with a BOW it is common to make efforts to take account for and exploit the fact that not all the words a document contains are equally representative of the text or the phenomena of interest within it. *Stop words* are lexical items that are extremely common in a given language or domain. Examples in English are *a*, *to*, and *would*—grammatical function words which are unlikely to encode useful information about a document. These are therefore often removed in pre-processing for NLP tasks. NLP libraries such as Natural
Language Toolkit \(^1\) and spaCy \(^2\) provide statistically generated lists of such stopwords, which I use to remove such terms from the debate transcripts. I also develop a list of procedural terms specific to parliamentary debates that are also removed in pre-processing for the experiments in Section 5.1. These include terms such as the Honourable Gentleman and give way, which occur frequently throughout parliamentary debates, but may not contribute to sentiment or topic information.

In order to select the lexical features that are likely to be most representative of the documents in a corpus, I apply term frequency-inverse document frequency (tf-idf) weighting to the items in the generated BOWs. Tf-idf is a statistical measure that scores and ranks each item by calculating its normalized term frequency (tf)—how often term \( t \) appears in document \( d \) divided by the number of terms in \( d \)—and the inverse document frequency (idf)—the logarithm of the number of documents \( n \) in the corpus divided by the number of documents that contain term \( t \):

\[
idf(t) = \log\left(\frac{n}{(df(t) + 1)}\right)
\]

Tf-idf is then calculated as the product of the term’s tf and its idf:

\[
tf-idf(t, d) = tf(t, d) \times idf(t)
\]

Features are then selected by choosing a threshold, and including in the text representation only terms which meet or exceed this.

As the standard approach to text representation, which is used in most of the studies reviewed in Chapter 3, I use BOW as a baseline approach in the experiments in Part II.

### 4.1.2 Word embeddings

In addition to the limitation described above, a further drawback of the BOW is its inability to handle out-of-vocabulary (OOV) items. With this model, any words that appear in the test set documents which are not also present in the

---

2. spaCy stop word list available at [https://spacy.io/usage/adding-languages#stop-words](https://spacy.io/usage/adding-languages#stop-words) accessed 13 July 2020.
vocabulary of the training set are unaccounted for, and cannot therefore be included as features. One way to mitigate this is to use word embeddings—vector representations of the contexts in which terms appear in a (large) corpus.

![Figure 4.2: Visual representation of a word embeddings vector space.](image)

Word embeddings place vector representations of terms that appear in similar contexts in the corpus in close proximity to one another in the embedding vector space. They can thus encode the fact that, for example, *student* and *pupil* represent similar concepts, and also that the relationship between *Paris* and *France* is similar to that between *Stockholm* and *Sweden* (see Figure 4.2). By converting the words in a document collection to embedding vectors, so long as they appear in the corpus from which the word embeddings were constructed, OOV items in the test set can be represented in the feature space.

Examples of such embeddings that have been pretrained on very large corpora are available. By looking up the words that appear in a dataset of interest (such as one of the parliamentary corpora used in this thesis), it is possible to convert them to embedding vectors for use as features to be passed to a classifier. There are a number of approaches to generating word embeddings.

**Distributional word embedding models**

Based on the *distributional hypothesis* [Harris, 1954], BOW-based approaches to word embedding generation include word2vec [Mikolov et al., 2013a]. For this, embeddings are constructed by training a neural network model to either predict the word most likely to appear in a given context (the continuous bag-of-words
(CBOW) model), or to predict the context of a target word (the skipgram model) (both Mikolov et al., 2013b). In order to better increase the ability of the topic identification system to capture tokens with similar meanings, I compare use of such static word embeddings with both a BOW approach and the contextual transformer-based embeddings described below in the multiclass classification experiments in Section 6.3.

A limitation of such distributional approaches is that they do not account for polysemy. That is, no matter the meaning of a word in different contexts, it is mapped to the same vector. For example, ‘bank’ would have the same vector when used to refer to both a financial institution and a geographical feature. Attempting to account for this, transformer-based word embedding models have been proposed, and have recently become popular (Rogers et al., 2020).

Transformer-based language models
A significant development in NLP in recent years has been the increase in the use of transformer-based word embedding models. The use of architectures such as Embeddings from Language Models (ELMo) (Peters et al., 2018) and Universal Language Model Fine-Tuning method (ULMFiT) (Howard and Ruder, 2018) has led to significant performance gains across a range of tasks. One of the reasons for this is that, unlike the distributional models described above, such models are able to take context into account by producing a range of different embedding vectors for each word. One of the most prominent transformer language models is BERT (bidirectional encoder representations from transformers) (Devlin et al., 2019), which has been used to obtain state-of-the-art performances on several datasets produced for sentiment polarity classification (e.g., Devlin et al., 2019; Sun et al., 2019b; Xu et al., 2019).

Large pre-trained BERT embedding models are available, which can be fine-tuned to fit data from the target domain and a specific task (such as document classification). These models have been trained on a large corpus of unlabelled text including Wikipedia (over 2,500 million words) and the Book Corpus (around 800 million words). Pre-trained BERT models are available at https://tfhub.dev/google/collections/bert/1, accessed 13 July 2020.

To produce the models, sentences from the corpora are input as an embedding layer in which each token is represented with three elements: a token embedding, a segment embedding, and a position embedding. These are fed to the transformer,
which consists of a number of encoder layers which train on two tasks: masked language modeling and next sentence prediction. As these tasks are performed bi-directionally, the resulting vectors capture contextual information from both before and after the target tokens, unlike the uni-directional training of ELMo and ULMFiT. The vectors that are produced in the final encoder layer while training on these tasks comprise the BERT embedding model. This model can then be fine-tuned by adding a task-specific layer, such as a sentiment classification layer. Figure 4.3 illustrates this process, where a special token \texttt{cls} is added to the input, and the output vector for this token is passed to the task-specific classifier as input. The maximum length of input documents is 512 tokens. The pre-trained BERT models are available in two sizes: base and large, and in cased (includes uppercase characters) and uncased (lowercased) versions.

![Diagram of BERT fine-tuning process](image)

Figure 4.3: A simplified representation of the process of fine-tuning a BERT model for classification.

Classifiers fine-tuned on BERT language models have achieved state-of-the-art results on a wide range of classification tasks, including those undertaken in this thesis: sentiment classification ([Devlin et al., 2019](#), [Rietzler et al., 2020](#), [Sun et al., 2019](#), [Tang et al., 2019](#), [Xu et al., 2019](#), and text classification ([Devlin](#))...
et al., 2019). However, performance of systems trained on such language models can vary greatly, is sensitive to domain changes, and does not always perform well on real-world tasks (Xia et al., 2020). I combine these models with fine-tuning for classification, and compare this approach with simpler BOW and static word embedding models in the experiments described in Part II.

4.2 Supervised machine learning methods

As parliamentary debate transcripts are complex, usually requiring expert human interpretation, the majority of the experiments presented in Part II rely on supervised methods. While such methods require labelled data, they are able to make use of the human knowledge of the world that is encoded in the class labels. As discussed above, supervised classification models take as input labelled representations of the data to be classified from which they learn to minimise a loss function so as to optimize classification performance. In Part II I evaluate supervised classification in both binary (for sentiment polarity) and multi-class (for opinion-topics/policy preferences) settings.

**Binary classification** A binary classifier takes as input a collection of \( l \) example documents \( \{x_i, y_i\} \) with \( d \) input features, where \( x_i \in \mathbb{R}^d \), and each class label \( y_i \in \{0, 1\} \) and outputs a predicted label (0 or 1) for each test example.

**Multi-class classification** The approach taken in this work for multiclass opinion-topic and policy preference identification is that of one-vs-all (sometimes referred to as one-vs-the-rest) classification. In this scenario, for each class label \( y_i \in \{y_0, y_1, ..., y_n\} \), binary classification is performed, with the two classes being \( \{y_i, \neg y_i\} \) (the target class or not the target class). Performance is evaluated by iterating over the set of class labels and aggregating results over classification on all the classes. This approach also allows instances to belong to more than one class (multi-label classification), as in Section 6.2.

In the rest of this section, I provide descriptions of the machine learning algorithms that I use for classification.
4.2.1 Support Vector Machines

Support vector machines (SVMs) are a machine learning method that attempts to maximise the margin between the two classes in either a linear or a high-dimensional feature space (Shalev-Shwartz and Ben-David 2014). This is achieved with the aid of support vectors, which pass through the datapoints nearest to the decision hyperplane, as shown in Figure 4.4.

![Figure 4.4: The maximum-margin hyperplane (thick, red line) of an SVM separating two classes. The example datapoints on the margins (dashed lines) are the support vectors.](image)

For SVMs, each label $y_i \in \{-1, 1\}$.

With $d$ features, hyperplanes in the feature space $\mathbb{R}^d$ consist of a vector $w$ and a constant $b$:

$$w \cdot x + b = 0 \quad (4.3)$$

The aim of the SVM is to find, for all training instances $\{x_i, y_i\}$, a hyperplane which is represented by the decision function:

$$f(x) = \text{sign}(w \cdot x + b) \quad (4.4)$$

with the parameter $w$ and the constant $b$. Here, the following two criteria must be satisfied:

4That is, negative labels are converted from 0 to $-1$. 
1. the value of \( f(x) \) should be equivalent to the label of the training instance:

\[
\text{sign}(w \cdot x_i + b) = y_i
\]  

(4.5)

2. for all datapoints the geometric margin (orthogonal distance from the hyperplane) should be at least 1:

\[
y_i(w \cdot x_i + b) \geq 1 \forall i
\]  

(4.6)

The distance \( d \) from the hyperplane to the closest data points on either side is maximised:

\[
d((w, b)x_i) = \frac{y_i(w \cdot x_i + b)}{||w||} > \frac{1}{||w||}
\]  

(4.7)

The maximal margin is obtained by minimising \( ||w|| \):

\[
\min_{w,b} \frac{1}{2}||w||^2
\]  

(4.8)

With a vector \( \alpha \) consisting of \( l \) Lagrange multipliers and a constant \( C \), this can be solved with the following equation:

\[
W(\alpha) = \text{argmin}(\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_iy_j\alpha_i\alpha_j(x_i \cdot x_j) - \sum_{i=1}^{l} \alpha_i)
\]  

(4.9)

under the following constraints:

\[
0 \leq \alpha_i \leq C \forall i
\]  

(4.10)

\[
\sum_{i=1}^{l} y_i\alpha_i = 0
\]  

(4.11)

In many cases, most of the resulting vectors \( \alpha_i \) have values of 0. Where \( \alpha_i > 0 \), these are considered to be support vectors, which define the optimal hyperplane. Lowering the value of the constant \( C \) results in a flexible hyperplane, allowing for misclassification of some of the data points whilst minimising the margin of error—that is, attempting to keep the value of \( y_i(x_i \cdot w + b) \) as close to 1 as possible.
SVMs are commonly used as a (strong) baseline both in sentiment analysis of legislative debate speeches (see Chapter 3), and in NLP more generally. I therefore employ SVM classification as a high baseline in the experiments in Chapters 5 and 6.

### 4.2.2 Artificial neural networks

Artificial neural networks (ANNs) consist of an input layer $X$ (the representation of the training examples, consisting of features $\{x_1, x_2, ..., x_n\}$), an output layer (the class labels, $Y$), and one or more ‘hidden’ layers of computational units (‘neurons’).

![Figure 4.5: Architecture of a neural network with $n$ input features $\{x_1, ..., x_n\}$, two hidden layers ($l$) of $n$ nodes $a_1$ to $a_n$, and an output layer that returns the predicted label $\hat{y}$.](image)

Each unit in the network is a function which takes as input the weighted sum of the outputs of the neurons from the previous layer that are connected to it, and forwards the output to those in the next layer. Networks with more than one hidden layer are considered to be ‘deep’ ANNs (Shalev-Shwartz and Ben-David, 2014).

With the dramatic growth in the size and availability of data in recent years,
ANNs have regularly obtained state-of-the-art performances on NLP tasks. However, as Chapter 3 shows, they have rarely been employed in the domain of parliamentary debates. I use the following neural network models for classification.

**Multi-layer Perceptron**

A multi-layer perceptron (MLP) is a basic ‘vanilla’ feed-forward neural network that has one or more hidden layers with any number of nodes in each layer. Each node in each layer is connected to every node in the following layer with weight \( w \), where \( w \in W \) (see Figure 4.5).

The network learns by adjusting parameters \( \theta \), which consist of the connection weights \( W \) and a constant \( b \), as each data instance is processed. This adjustment is performed based on calculation of the loss \( L \).

Each hidden layer \( l \) receives the inputs \( X \), and each node \( i \) in that layer calculates

\[
z_i^{[l]} = w_i^{[l]^T} x + b_i^{[l]}.
\]

(4.12)

The result is then output to the next layer via a non-linear activation function, \( a \) as \( a_i^{[l]}(z_i^{[l]}) \). In the final layer, a linear activation function outputs either 1 or 0 as the class label prediction \( \hat{y} \). With \( m \) data points and the cost function

\[
J(W, b) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}, y),
\]

(4.13)

the gradient with respect to \( W \) is then calculated using gradient descent by back-propagation, and the weights are adjusted accordingly. An optimization algorithm is used to find the parameters \( \theta \) that minimize the cost function.

I use MLPs as simple neural models for the classification experiments in Sections 5.1, 5.2, 6.3 and 7.1.

**Convolutional Neural Network**

A Convolutional Neural Network (CNN) is a regularised form of ANN that uses convolution rather than general matrix multiplication in at least one of its hidden layers (Goodfellow et al., 2016). Convolution is an operation on two functions of a real-valued argument that produces a third function expressing how the shape of one is modified by the other. The convolution of functions \( f \) and \( g \) (where,
for text input, $t$ may be a token in a sequence) is defined as the integral of the product of the two functions at timestep $t$ after one is reversed:

$$(f * g)(t) = \int_{0}^{\infty} f(\tau)g(t - \tau)d\tau \quad (4.14)$$

In an ANN, a typical convolutional layer actually consists of several sub-layers, as illustrated in Figure 4.6. To perform convolution, a sliding window passes over the matrix of input vectors convolving the values in the window with a filter. The result of this convolution passes to the activation function. The output of this is then passed to a pooling layer, which serves to reduce the size of the representation, and retain only the more salient features in the layer. The final hidden layer in the network is a fully connected layer with a linear activation function which outputs the predicted label.

![Figure 4.6: Architecture of a convolutional layer in a CNN.](image)

The feature maps generated by CNNs are somewhat equivalent to n-grams of the chosen window size, adding similar contextual information to the representation. They are computationally efficient and have been shown to perform competitively on many NLP tasks (Dauphin et al., 2017), including sentiment classification (Johnson and Zhang, 2015b; Zhang et al., 2015). I explore the use of CNNs for classification tasks in Sections 6.3 and 7.1.

**Activation functions** There are a range of options for activating the output of the nodes in a neural network. For hidden layers, I use rectified linear units (ReLUs), which are considered to be the default choice due to their speed and efficiency (Goodfellow et al., 2016). The ReLU function is defined as:

$$g(x) = \max\{0, x\}$$

such that the output is equal to the input where $x > 0$, and zero otherwise.
For classification, it is necessary to use a linear function in the output layer in order to produce a discrete value for the predicted label $\hat{y} \in [0, 1]$. I use the sigmoid function $S$, which is again considered to be the default choice. $S$ is defined as:

$$S(x) = \frac{1}{1 + e^{-x}}$$  \hspace{1cm} (4.16)

**Optimisation** Another decision to be made when using an ANN is the choice of algorithm used to optimize the parameters that minimize the loss. In Section 5.1, based on the results of initial experiments, I use the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) to optimize the parameters of the MLP. For all other experiments involving ANNs, I use adaptive moment estimation (Adam) (Kingma and Ba, 2015), which is considered to be robust to other parameter choices. While a full description is outside the scope of this thesis, an overview of optimisation algorithms, see Chapter 8 of Goodfellow et al. (2016).

### 4.2.3 Multi-task learning

Observing that it can be ‘easier to learn several hard tasks at one time than to learn these same tasks separately’, Caruana (1993) proposed *Multi-task learning*. This is a form of inductive transfer, where, in learning a hypothesis (such as a classifier) for one task, an inductive bias is provided by a second task (and vice-versa), causing the model to prefer hypotheses that fit both. In Chapter 7, in which I seek to obtain predicted labels for two tasks given the same input, I investigate the efficacy of multi-task learning. While numerous approaches have been taken to multi-task learning (for an overview, see Ruder (2017)), for these experiments, in which the two tasks (1) are closely related (both are classification), and (2) share the same input, I use a hard parameter-sharing approach. In this setting, the network consists of one or more shared hidden layers, the output of which is passed to the separate task-specific layers. For further details of this architecture, see Section 7.1.2.
4.3 Unsupervised topic modelling

In the initial topic identification experiments of Section 6.1.2, I explore the utility of identifying the subjects under discussion in parliamentary debates using topic modelling. For this, I use arguably the most well-known topic modelling method, **LDA** (Blei et al., 2002). LDA is a generative model of word counts. The model calculates document-topic and topic-word distributions to generate the documents. Here, a topic is defined as a mixture over the words in the vocabulary where each word has a probability of belonging to each topic. A document is a mixture over the topics, such that each document can be composed of multiple topics. For a particular document, the topic proportions across all topics sum to 1, and for a given topic, the word probabilities also add to 1. The outputs are a set of \( k \) topics (a parameter to be set by the user), and the probabilities of each word in the vocabulary being generated under each topic. Topics are then typically evaluated manually by observation of the \( n \) most probable words for each topic (with \( n \) also selected by the user).

4.4 Performance metrics

A number of performance metrics are commonly used to evaluate the performance of NLP systems. Depending on the aims of the task and the distribution of the data, these may put more emphasis on **true positives** (examples correctly labelled as positive), **false positives** (examples incorrectly labelled as positive), **true negatives** (correctly labelled negatives, and **false negatives** (incorrectly labelled negatives). I use the following metrics to evaluate the performance of the machine learning classification systems.

**Accuracy**

*Accuracy* is the percentage of correct classifications—true positives and true negatives—of all predictions made. In scenarios in which the classes are fairly evenly balanced, this can be a straightforward and easily interpreted measure. However, for cases in which the datasets are unbalanced (for example, where one class accounts for most of the examples), accuracy may provide uninformative or misleading information. In such a situation, a classifier that simply predicts the majority class for all examples would produce very high accuracy scores, but this
would provide no information about the ability of the system to learn from the data.

**F1-score**

The *F1-score* is designed to provide balance between a classifier’s precision (*true positives / true positives + false positives*) and recall (*true positives / total positive true labels*). It can therefore account for class imbalances in the data. F1 is calculated as:

\[
2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

In multi-class settings, precision, recall, and F1 can be calculated using either macro- or micro-averaging. For the former, the metric is computed independently for each class, and the reported score is the average over all classes. Here, majority classes may dominate the metric, and performance on the less-well represented classes is marginalised. For micro-averaging, all classes are weighted according to the number of (true-labelled) examples they contain, hence providing a better measure of a classifier’s performance over all the classes. In the micro-averaged setting, precision, recall, and F1 are equal.

I report F1 alongside accuracy in all the experiments in Part II, and both macro- and micro-averaged F1 for multi-class classification in Chapters 6 and 7.

### 4.5 Inter-annotator agreement measures

The corpora created for the experiments in Sections 5.1, 6.3, and 7.1 include manually applied class labels produced by multiple human annotators. In order to measure the validity of these labels, I calculate inter-annotator agreement rates using the following commonly used measures.

**Cohen’s kappa**

For comparison of two sets of labels, I report Cohen’s *kappa* (*κ*), which is calculated as:

\[
\kappa = 1 - \frac{1 - p_o}{1 - p_e}
\]

(4.18)
where $p_o$ is the percentage of examples on which the annotators agree, and $p_e$ is the probability of chance agreement. The latter is based on the chi-square matrix and is calculated as:

$$p_e = \frac{(cm^1 \cdot rm^1)}{n} + \frac{(cm^2 \cdot rm^2)}{n} \tag{4.19}$$

where $cm^1$ is the column 1 marginal (total observations in the column), $cm^2$ is the column 2 marginal, $rm^1$ is the row 1 marginal, $rm^2$ is the row 2 marginal, and $n$ represents the total number of observations.

The $\kappa$ score is a number between 0 and 1 inclusive, with higher values representing greater agreement. It is commonly interpreted as follows: values $\leq 0$ indicate no agreement, $0.01 - 0.20$ represent no to slight agreement, $0.21 - 0.40$ is fair, $0.41 - 0.60$ moderate, $0.61 - 0.80$ substantial, and $0.81 - 1.00$ as almost perfect agreement (Cohen, 1960).

**Fleiss’ kappa**

Fleiss’ $\kappa$ is used for measurement of labels provided by three (or more) annotators, and is used to validate the manually applied labels produced for the experiments in Sections 5.1 and 6.2. It is calculated as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{4.20}$$

where $\bar{P} - \bar{P}_e$ is the degree of agreement actually achieved above chance, and $1 - \bar{P}_e$ is the degree of agreement above chance that is actually attainable.

With the $N$ annotators, the number of ratings per annotator $n$, and the number of classes $k$, to obtain $\bar{P}$ and $\bar{P}_e$, we must calculate the proportion of all assignments which were made to the $j$-th class, $p_j$, and the proportion to which raters agree for the $i$-th example, $P_i$:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \quad , \quad \bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{4.21}$$

where

$$P_i = \frac{1}{n(n-1)} \left[ \left( \sum_{j=1}^{k} n_{ij}^2 \right) - (n) \right] \tag{4.22}$$
and

\[ p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \]  

(4.23)

For this measure, \( \kappa \) can be interpreted as: < 0: poor; 0.01-0.20: slight; 0.21-0.40: fair; 0.41-0.60: moderate agreement, 0.61-0.80: substantial agreement, and 0.81-1.00: near perfect agreement (Landis and Koch [1977]).

### 4.6 Chapter summary

In this chapter, I have described the approaches to text representation and feature selection, machine learning methods, performance metrics, and statistical annotator agreement measures, which I apply, and adapt where necessary, for the experiments in Part II.
Part II

Experiments and analysis
Chapter 5

Sentiment classification

‘Virtually every term of political discourse has two meanings. One is its literal meaning, and the second, which is often quite different, is its usage in political discourse.’

Noam Chomsky, linguist

In this chapter, I present analysis of the sentiment expressed by speakers towards the motions debated in the House of Commons. In these experiments, as in previous work (e.g., Burford et al. 2015, Onyimadu et al. 2013, Salah 2014, Thomas et al. 2006), the topics or policies to which these relate are unknown. For each debate, I attempt to ascertain whether or not each speaker positions themself in support of, or in opposition to, the speaker who tables the motion of the debate. This differs from approaches to argument mining in which a key part of the problem may be to determine whether a statement is related to the target at all (see, for example, Menini et al. 2018).

I present two sets of experiments conducted on two different datasets constructed for this purpose. Firstly, I compare the use of speakers’ division votes with manually annotated polarity labels. I also introduce a two-step, motion-dependent sentiment analysis model of parliamentary debates in which the sentiment of both speeches and motions are classified. For this model, I 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, I evaluate the use of n-gram textual features and a range of contextual features extracted from metadata related to the speakers.
CHAPTER 5. SENTIMENT CLASSIFICATION

Secondly, using a larger dataset, I compare the performance of state-of-the-art machine learning approaches and different language models for this task.

5.1 Motion-dependent sentiment analysis with vote-derived and manual class labels

In this section, I examine two aspects of the task of sentiment analysis of parliamentary debate transcripts: the use of different class labels for supervised classification and the effect of the semantic structure of the debates on speech sentiment polarity. The experiments I present here test hypothesis H1: MPs’ votes are unreliable sentiment/stance class labels, as they do not always reflect the opinions and positions expressed in their speeches, and H2: The polarity shifts caused by the discourse structure of parliamentary debates can be mitigated by applying a two-step, debate motion-dependent classification model.

Parts of this section have previously been published in Abercrombie and Batista-Navarro (2018a) and Abercrombie and Batista-Navarro (2018b).

5.1.1 Sentiment polarity class labels

Previously existing datasets for sentiment analysis of legislative debates tend to rely on speakers’ votes as sentiment polarity labels (e.g., Salah, 2014; Thomas et al., 2006). However, it is widely recognised that MPs are, to a large extent, constrained in their voting behaviour, as they are often under pressure to vote along party lines irrespective of their personal opinions (Crewe, 2015; Norton, 1997; Searing, 1994). With their speeches, on the other hand, MPs often communicate to their constituents or the wider public in ways that do not necessarily toe the party line. Indeed, analysis in the field of political science provides evidence for this disparity between politicians’ speeches and voting behaviour (Schwarz et al., 2017). Proksch and Slapin (2015) suggest that MPs in the UK Parliament often use their speeches to communicate to their constituents, and that both they and their parties use debates to demonstrate opposition to their opponents—while ultimately voting differently to the way these signals might suggest. For instance, in Example 5.1 the speaker appears to be opposed to the motion, yet has been recorded as voting in support of it:
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).

In addition to such political machinations, there are a number of other factors which call into question the reliability of votes as proxies for MP sentiment. MPs may change their mind between speech and vote (as suggested by Salah (2014)). They are even known to vote erroneously on occasion, as described by Paul Flynn, MP Flynn (2012) and Caroline Lucas, MP Lucas. Additionally, those MPs acting as tellers have their votes counted as being for the side on which they are counting, irrespective of their position on the motion in question.

All of this means that, while convenient to obtain, votes may not be accurate reflections of the ground-truth opinions expressed in the content of MPs’ debate speeches. An alternative form of class labelling may be required for effective sentiment classification using supervised machine learning methods. For these experiments, I therefore constructed a dataset with two distinct sets of class labels for comparison: vote-derived labels and manually annotated labels.

5.1.2 Semantic Structure of House of Commons Debates

The following two characteristics of debate structure are pertinent to the sentiment detection task:

Motion sentiment Speaker sentiment is present not only in debate speeches, but also in motions, where it may also be positive or negative. In proposing a motion, an MP expresses sentiment towards the policy, piece of legislation, or state of affairs in question. This differs from traditional sentiment analysis

---

domains, where the target of any sentiment tends to be a neutral entity such as a product or service.

Annotators are able to label motions as being either positive or negative with high agreement (see Subsection 5.1.3), and these labels correlate highly with the party affiliations of the MPs that propose the motions (Subsection 5.1.4).

**Double-negative effect**  As noted by Thomas et al. (2006), the language used in legislative debates to express positive or negative speech sentiment is radically altered depending on the sentiment polarity of its target—the Bill or motion under debate. A form of polarity shifting can occur, whereby speakers may use typically negative language to demonstrate positive sentiment and vice versa. For example, if a motion commends the actions of the Government, speeches in support of the motion are likely to incorporate positive language, while those opposing it will tend to be characterised by negative language. If, however, a motion condemns Government policy, supportive speeches are themselves also likely to contain typically negative language, and opposing ones positive language. For instance, Example 5.2 provides an instance of a negative motion (proposing that legislation be *annulled*), with a speech which uses negative language to support it:

**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 country 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?* (5.2)

This ‘double negative’ effect presents potential 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.

Based on these observations, I proposed a two-stage, motion-dependent sentiment classification model, in which opinions expressed in both motions and
speeches are analysed.

For this reason, I included in the dataset 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 or position, I also included a set of motion sentiment labels derived from the party affiliations of the MPs who propose the motions: *positive* if they are members of the governing party or coalition, *negative* if not.

**Research objectives** The experiments in this section are therefore focused on the following two objectives:

1. To assess the validity of class labels derived from speakers’ votes, and the extent to which they are reflective of ground-truth sentiment polarity.

2. Evaluate the use of the proposed two-step, motion-dependent debate model and Government/Opposition motion model to mitigate the polarity shifting effects of motion sentiment when conducting sentiment analysis of debate speeches.

5.1.3 Corpus construction

**Data collection**

I used the HanDeSet corpus (see Section 2.4), which consists of records of House of Commons debates from May 1997 to July 2017 (when I began this part of the project). This period was chosen in order to obtain a sufficient quantity of speeches for which there exist associated division votes, and because the record for these years is most complete in terms of metadata (see Section 2.2). Each file contains transcripts of a full days’ activities in the House of Commons: a number of debates, questions, and other parliamentary affairs, such as prayers and the reading of public petitions. I automatically selected debates under ‘major-heading’ elements in the XML files as these are debates which often culminate in divisions (votes). I retained only debates that contain a motion and precisely one division, under the assumption that in such instances each member’s vote represents their sentiment towards the motion under debate. I manually excluded general motions of the type ‘This House has considered …’ (see Chapter 2) to leave only debates with substantive motions, as, by their nature, these demand polarised stances to be taken by MPs.
Data processing  Parliamentary speeches incorporate much set, formulaic discourse related to the operational procedures of the chamber, which I automatically removed as such content does not concern the motion or the speakers’ opinions towards it. This includes 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. Additionally, I 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) and Thomas et al. (2006), I considered a member’s speech to be the concatenated content of all their utterances (individual speech segments or paragraphs) in a given debate. For comparison of manual and vote labelling methods, I retained all speeches made by MPs who appear in the division of the given debate along with a record of their vote. I omitted speeches made by the member of the assembly that proposes the motion, as, by definition, they speak in support of the proposal.

Also following Salah (2014), I removed speeches totalling fewer than 50 words. In order to facilitate manual labelling, I restricted the quantity of text to be read by human annotators by including only those speeches comprised of five utterances or fewer. Finally, I removed quotations within speeches. These can reflect opposing or different points of view to those of the speaker (where, for example, an opponent is quoted), and could therefore represent confounding features for the learning of a classifier.

Annotation

In addition to myself, I recruited two further annotators. These were both first language English speakers, university graduates, UK citizens, and self-reported as being familiar with British politics and the UK parliament. Following the procedure of Pustejovsky and Stubbs (2012), we developed a set of annotation guidelines in a two-round cycle and produced annotations for a randomly selected subsection (250 speech examples, 20 per cent) of the speeches.

---

2I automatically removed the following procedural language: names of MPs mentioned in speeches (which are inserted by the reporters), utterances solely concerned with ‘giving way’ or making interventions, utterances concerning points of order.

3The annotation guidelines are available in Appendix A.
Manual sentiment labelling was carried out on this subsection in two rounds of the following cycle: randomly selected subsection of the speeches.

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

1. Production/update of annotation guidelines.

2. Two annotators labelled the corpus subsection (annotators 1 and 2 for round 1, annotators 1 and 3 for round 2).

3. Inter-annotator agreement calculated, and disagreement analysis performed.

**Annotator agreement** To assess the validity of the manually applied labels, I calculated Cohen’s kappa ($\kappa$) after each round of annotations. I 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.

<table>
<thead>
<tr>
<th>Round</th>
<th>Annotation guidelines used</th>
<th>Motion $kappa$</th>
<th>Speech $kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Version 1 (annotators 1 &amp; 2)</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>2</td>
<td>Version 2 (annotators 1 &amp; 3)</td>
<td>0.91</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 5.1: Inter-annotator agreement (Cohen’s kappa) for motion and speech sentiment polarity labels following the first and second versions of the annotation guidelines.

Inter-annotator agreement on the first round of annotation was found to be ‘moderate’\(^4\) 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 either in favour of, or against something.

I examined the cases on which the annotators disagreed, manually identifying the probable causes of disagreement presented in Table 5.2. To address the issues raised by this analysis, I updated the annotation guidelines, clarifying the instructions and adding further example cases. In particular, I defined a protocol for handling motions which call on the Government for action, but which

\(^4\)Following the guidelines for interpretation of $\kappa$ of Landis and Koch (1977).
CHAPTER 5. SENTIMENT CLASSIFICATION

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

Table 5.2: Causes of disagreement for the first round of annotation according to manual identification.

can be seen as attacking its position. These are common in the corpus and had accounted for 85 per cent of annotator disagreement on motion sentiment. I also provided the annotators with more contextual information, by adding the MPs’ party affiliations to the data they were presented with.

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.\(^5\) 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.

I 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 I identified some as being either off-topic or highly ambiguous, as in Example 5.3:

\(^5\)Note that annotator 2 was replaced by annotator 3 for the second round of annotation due to unavailability.
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.

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 (for example, the UK’s membership of the EU), the speaker’s sentiment towards the motion is likely to seem ambiguous. As this information is not necessarily present in the debate in question, annotators may be forced to rely on their knowledge or assumptions about speakers. The presence of such speeches in House of Commons debates makes it unlikely that 100 per cent agreement could be achieved on this task without further contextual clues.

The final version of the annotation guidelines are available in Appendix A. Following the guidelines, the job of the annotator can be summarised as follows:

1. 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.

2. 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’).

Following this process, as the principal annotator, I then followed the final guidelines to produce the gold standard labels for the remainder of the corpus.
Corpus Description The corpus is available at Mendeley Data. It consists of 1,251 motion-speech units taken from 129 separate debates, with each unit consisting of 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 1,049 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 two sentiment labels: one manually applied, as described above, and a Government/opposition label obtained automatically according to the party affiliation of the MP proposing the motion. 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, and speaker name.

Some prior work omitted debates on which less than 20 per cent of speakers had voted for the losing side in order to avoid the inclusion of one-sided debates (Thomas et al., 2006). In the House of Commons, recorded divisions only take place in cases in which the vote cannot be easily determined by an oral vote, so this step was unnecessary.

<table>
<thead>
<tr>
<th>Target</th>
<th>Label type</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion</td>
<td>Government/opposition</td>
<td>71 (55.0%)</td>
<td>58 (45.0%)</td>
</tr>
<tr>
<td></td>
<td>Manually annotated</td>
<td>67 (51.9%)</td>
<td>62 (48.1%)</td>
</tr>
<tr>
<td>Speech</td>
<td>Vote-derived</td>
<td>713 (57.0%)</td>
<td>537 (43.0%)</td>
</tr>
<tr>
<td></td>
<td>Manually annotated</td>
<td>702 (56.5%)</td>
<td>544 (43.5%)</td>
</tr>
</tbody>
</table>

Table 5.3: Occurrences of sentiment labels in the HanDeSet corpus.

In the corpus, manually applied motion labels are approximately evenly balanced; the vote-derived labels are slightly skewed towards the positive class, as can be seen in Table 5.3. Concurrence between the vote labels and manually annotated labels is 92.8 per cent. It appears that, while the majority of speeches...
reflect the voting behaviour of the speaker, some do not, and that division votes may not therefore be entirely reliable as sentiment polarity labels.

In general, MPs both speak and vote along party lines, as shown in Figure 5.1. In this dataset, the smaller parties and the SNP have always voted as a block and were assigned the same manual sentiment label in every debate. The major UK-wide parties exhibit rather more rebellion, for both speech sentiment (Labour Party (Lab): 8.2%, Conservative Party (Con): 6.6%, Liberal Democrats (LD): 2.8%) and vote (Lab: 4.2%, Con: 1.1%, LD: 0.0%).

Upon examination of these ‘rebel’ speeches, I found 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) ethical matters of ‘moral conscience’ for which the whip is not usually employed (Crewe 2015), such as stem cell research, or (c) issues that are known to divide parties, such as membership of the European Union (EU).

In several of the speeches, a speaker stated explicitly that they would vote one way, only to have their vote recorded for the opposing side, confirming the potential unreliability of votes as ground-truth label for speaker sentiment.

### 5.1.4 Debate speech sentiment models

As noted in Subsection 5.1.2, the motions tabled in parliamentary debates express either positive or negative sentiment towards their targets, and members of the chamber speak either in support of, or in opposition to the motion. I therefore

![Figure 5.1: Tree diagrams showing the number of example speeches by party affiliation, and how many of the associated class labels represent ‘rebel’ votes and speeches (inset) in which the sentiment labels support and oppose the majority of the speaker’s own party, respectively.](image-url)
proposed a two-stage sentiment analysis model in which the sentiment of each motion towards the (unknown) subject of the debate is first determined, before performing sentiment analysis on the corresponding speeches. I compared the performance of the following three models (illustrated in Figure 5.2):

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

- **Model 2**: A two-step, motion-dependent *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, I also compared two methods of applying sentiment labels to the motions:

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

2b: Under the assumption that motions proposed by the Government are positive, and those proposed by other parties are negative, debates are labelled and classified by the party affiliation of the MP that proposes the motion—positive if they are a member of the governing party or coalition, negative if not.

5.1.5 Experiments

I performed experiments to compare sentiment classification performance using systems that comprise combinations of the following:

- Two machine learning methods\(^8\)
  - SVM: support vector classification with a linear kernel.
  - MLP: a neural network with one hidden layer of size 100, using ReLU activation, sigmoid activation in the output layer, \( \text{L-BFGS} \) 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 motion-independent, one-step speech sentiment model, and the motion-dependent, two-step motion-speech sentiment model. For the motion-speech model, I also compared 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.

\(^8\)I used the implementations from \url{https://scikit-learn.org} (accessed 12 December 2020), with default settings.
• The following input 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. 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, I also followed 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, I added that Member’s party (at the time of the debate in question) as a further contextual feature.

  These features were vectorized and concatenated in the combinations presented in subsection 5.1.6.

5.1.6 Results

I present the results of speech sentiment classification using 10-fold cross-validation in Table 5.4. Due to the slight imbalances in positive and negative class labels, F1 scores are reported in addition to accuracy. The highest accuracy and F1 scores overall, using both labelling methods, are consistently achieved using all features to train the MLP classifier.

For classification of the motions, 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 $\kappa = 0.91$), this is perhaps close to the optimal performance that might 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
Table 5.4: Speech sentiment classification results: accuracy and F1 scores for combinations of debate model, motion and speech label types, and features (speech text, party, debate ID, and motion text) using SVM and MLP classifiers. Highest overall scores for each metric are highlighted in bold, best scores using text features only are underlined.

<table>
<thead>
<tr>
<th>Debate model</th>
<th>Motion label</th>
<th>Speech label</th>
<th>Features</th>
<th>SVM</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acc.</td>
<td>F1</td>
</tr>
<tr>
<td>Vote</td>
<td>Text only</td>
<td></td>
<td>Text only</td>
<td>64.3</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>Text, Party</td>
<td></td>
<td>Text, Party</td>
<td>78.8</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>Text, Party, ID</td>
<td></td>
<td>Text, Party, ID</td>
<td>82.7</td>
<td>0.848</td>
</tr>
<tr>
<td></td>
<td>Text, Party, ID, Motion</td>
<td></td>
<td>Text, Party, ID, Motion</td>
<td>82.6</td>
<td>0.847</td>
</tr>
<tr>
<td></td>
<td>Party, ID</td>
<td></td>
<td>Party, ID</td>
<td>83.3</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td>Party, ID, Motion</td>
<td></td>
<td>Party, ID, Motion</td>
<td>83.5</td>
<td>0.854</td>
</tr>
<tr>
<td></td>
<td>Text only</td>
<td></td>
<td>Text only</td>
<td>66.7</td>
<td>0.718</td>
</tr>
<tr>
<td></td>
<td>Text + Party</td>
<td></td>
<td>Text + Party</td>
<td>76.2</td>
<td>0.791</td>
</tr>
<tr>
<td>Manual</td>
<td>Text, Party, ID</td>
<td></td>
<td>Text, Party, ID</td>
<td>79.7</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>Text, Party, ID, Motion</td>
<td></td>
<td>Text, Party, ID, Motion</td>
<td>79.8</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>Party + ID</td>
<td></td>
<td>Party + ID</td>
<td>79.9</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>Party, ID, Motion</td>
<td></td>
<td>Party, ID, Motion</td>
<td>80.0</td>
<td>0.822</td>
</tr>
<tr>
<td>Vote</td>
<td>Text only</td>
<td></td>
<td>Text only</td>
<td>72.9</td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td>Text, Party</td>
<td></td>
<td>Text, Party</td>
<td>83.9</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>Text, ID, Party</td>
<td></td>
<td>Text, ID, Party</td>
<td>86.1</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td>Text, ID, Party, M</td>
<td></td>
<td>Text, ID, Party, M</td>
<td>86.5</td>
<td>0.859</td>
</tr>
<tr>
<td></td>
<td>Party, ID</td>
<td></td>
<td>Party, ID</td>
<td>83.3</td>
<td>0.821</td>
</tr>
<tr>
<td></td>
<td>ID, Party, Motion</td>
<td></td>
<td>ID, Party, Motion</td>
<td>83.2</td>
<td>0.818</td>
</tr>
<tr>
<td>2a: Motion Classifier</td>
<td>-Speech</td>
<td>Text only</td>
<td>Text only</td>
<td>74.7</td>
<td>0.710</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text, Party</td>
<td>Text, Party</td>
<td>81.0</td>
<td>0.760</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text, Party, ID</td>
<td>Text, Party, ID</td>
<td>83.1</td>
<td>0.794</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text, Party, ID, Motion</td>
<td>Text, Party, ID, Motion</td>
<td>82.9</td>
<td>0.790</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Party, ID</td>
<td>Party, ID</td>
<td>80.7</td>
<td>0.747</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Party, ID, Motion</td>
<td>Party, ID, Motion</td>
<td>79.6</td>
<td>0.734</td>
</tr>
<tr>
<td>Vote</td>
<td>Text only</td>
<td></td>
<td>Text only</td>
<td>73.1</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>Text, Party</td>
<td></td>
<td>Text, Party</td>
<td>85.1</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td>Text, Party, ID</td>
<td></td>
<td>Text, Party, ID</td>
<td>87.5</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td>Text, Party, ID, Motion</td>
<td></td>
<td>Text, Party, ID, Motion</td>
<td>87.8</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>Party, ID</td>
<td></td>
<td>Party, ID</td>
<td>82.9</td>
<td>0.820</td>
</tr>
<tr>
<td></td>
<td>Party, ID, Motion</td>
<td></td>
<td>Party, ID, Motion</td>
<td>84.9</td>
<td>0.848</td>
</tr>
<tr>
<td>2b: Motion Government</td>
<td>-Speech</td>
<td>Text only</td>
<td>Text only</td>
<td>74.3</td>
<td>0.736</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text, Party</td>
<td>Text, Party</td>
<td>72.6</td>
<td>0.799</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text, Party, ID</td>
<td>Text, Party, ID</td>
<td>84.8</td>
<td>0.828</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text, Party, ID, Motion</td>
<td>Text, Party, ID, Motion</td>
<td>84.4</td>
<td>0.824</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Party, ID</td>
<td>Party, ID</td>
<td>80.8</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Party, ID, Motion</td>
<td>Party, ID, Motion</td>
<td>80.6</td>
<td>0.770</td>
</tr>
</tbody>
</table>

are brought by the Government in an effort to pass laws. Many negative motions include structures such as ‘(this House) believes that/noted that/disagrees
Table 5.5: Top 10 most discriminating positive and negative n-gram features for motion sentiment classification, ranked by learned coefficients of the SVM classifier using manually annotated labels.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>security</td>
<td>notes</td>
</tr>
<tr>
<td>2</td>
<td>connection</td>
<td>amend</td>
</tr>
<tr>
<td>3</td>
<td>given</td>
<td>believes</td>
</tr>
<tr>
<td>4</td>
<td>purposes</td>
<td>calls</td>
</tr>
<tr>
<td>5</td>
<td>general</td>
<td>government</td>
</tr>
<tr>
<td>6</td>
<td>new</td>
<td>calls government</td>
</tr>
<tr>
<td>7</td>
<td>schedule</td>
<td>dated</td>
</tr>
<tr>
<td>8</td>
<td>proceedings</td>
<td>eu</td>
</tr>
<tr>
<td>9</td>
<td>session</td>
<td>disagrees</td>
</tr>
<tr>
<td>10</td>
<td>programme</td>
<td>number</td>
</tr>
</tbody>
</table>

with/calls on the Government to...’. This is also reflected in the most discriminating n-gram features (see Table 5.5), which suggests that motion polarity can often be predicted based on these verbs, which appear in the head position in the motions.

5.1.7 Discussion

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.

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 exclusively textual features are considered, while with division vote labels, the inclusion (or exclusive use) of metadata leads to considerable gains in performance (see Figure 5.3).

It therefore appears that textual features correlate more closely with 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
Figure 5.3: Comparison of results using vote-derived and manually annotated class labels (using MLP classifier) in two conditions: with contextual metadata features excluded or included.

...more indicative of speech sentiment (H1), considering the associated costs and the relatively small differences in performance, use of division votes may be the more pragmatic choice for this task for practical purposes.

Debate Models

Compared to the motion-independent speech model, use of the motion-dependent 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 (H2).

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.

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 almost as effective as training a classifier on manually annotated labels (model 2a). This suggests that a motion-dependent model can be used without the need for costly...
manual annotations, at least as far as motion sentiment labels are concerned.

Features

Textual Features

For textual features, the inclusion of bi- and trigrams does not appear to significantly improve speech classification performance over the use of only unigrams for this task, particularly for the two-step models (see Figure 5.4).

Figure 5.4: 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 coefficients also reveals that few bigrams and no trigrams feature in the top 10 most discriminating features (see Table 5.6). Examination of these predictive items underlines the fact that discriminating textual features for this task are not, on the whole, those that might 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 I found 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’, avoiding negative language in these debates, whatever point they may be making.

![Table 5.6: 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.](image)

The fact that terms that might normally be objective can indicate positivity or negativity in this domain, 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 having 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 terms such as ‘fox’ or ‘Wales’ become indicative of positive and negative sentiment polarity respectively.

I also found that an equal number of positive and negative scoring lexical items are indicative of the corresponding and the opposing sentiment polarity—that is, 30% of the discriminating features that have either mean positive or mean negative scores in SentiWordNet are used to express that same sentiment, while

---

30 per cent are used to express the opposite sentiment. Notably, among these appear the names (or initials) of three of the largest political parties, with SNP indicating positive, and Lab or ‘Labour party’, and ‘Conservative’ or ‘Tory’ (a colloquial term for a member of the Conservative party) indicating negative sentiment.

**Contextual Metadata Features**

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 in which MPs do not vote along party lines (as discussed in Section 5.1.3). In such cases, party affiliation can be a confounding feature and lead to incorrect classification.

**Classifiers**

Using textual features only, there is little difference between the performance of the two classifiers. 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.

**Error Analysis**

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

I 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. I observed the following:

1. **Speech length** Incorrectly classified speeches tended to be longer than average for the dataset, with a mean length of 218.8 tokens, compared to 167.8 words for the whole corpus.

2. **Party affiliation** Inclusion of this feature may have been a confounding variable for one of the following reasons:

   (a) Speech sentiment labels did not agree with the majority of that speaker’s party (19 per cent of errors).
(b) The speaker’s party was split in the debate concerned (11.9 per cent).

(c) The speaker was the only member of their party in the debate in question (22.4 per cent).

(d) The debate featured only one speech (and therefore one party) (4.5 per cent).

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

When using only textual features as input, I also examined examples in which the best performing (highest accuracy) configuration—the motion-speech model with SVM and manually annotated 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 5.4, in which the speaker talks about a different event than the target of the motion).

**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? (5.4)

Even when speeches do contain subjective language directed at the motion, multiple opinion targets, such as other MPs, parties, and topics, can also be present, complicating the task of sentiment classification. For instance, in Example 5.5, the speech extract includes five different target entities.
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.

5.1.8 Conclusion

The results presented in this section 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.

The motion-dependent Motion-speech model outperforms a simple motion-independent 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, I 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.
5.2 Sentiment classification using transformer-based language models

As the results of Section 5.1 demonstrate, there is little difference between classification performance using manually annotated and vote-derived sentiment polarity class labels. Considering the costs associated with producing the former, and the resulting limited sizes of the available datasets for sentiment analysis in this domain, I constructed another dataset that is sufficiently large to test a range of state-of-the-art text representation and machine learning methods.

Parts of this section have previously been published in Abercrombie and Batista-Navarro (2020a).

Sentiment polarity classification at scale While previously existing corpora for sentiment analysis of parliamentary debates are rather small (see Table 5.7), datasets typically used in other domains (such as product reviews or social media) can run into the hundreds of thousands of labelled examples. As state-of-the-art ANN methods using contextual word embeddings tend to benefit from larger datasets (Ng, 2019), there is a need to develop more extensive corpora for this task.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size (speeches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConVote</td>
<td>Thomas et al. (2006)</td>
</tr>
<tr>
<td>HanDeSet</td>
<td>Section 5.1</td>
</tr>
<tr>
<td>ParlVote</td>
<td>This section</td>
</tr>
</tbody>
</table>

Table 5.7: Size in number of example speeches of publicly available datasets for supervised speech-level sentiment analysis of legislative debate transcripts.

Research objectives With the experiments presented in this section, I aim to fulfil the following objectives:

1. Exploiting the fact that it is possible to collect a large vote-labelled dataset, and that deep learning methods tend to benefit from large amounts of training data, evaluate how sentiment polarity classification performance is affected by increases in the scale of parliamentary datasets.
2. Considering their successes in other domains, assess the performance of state-of-the-art transformer-based contextual embedding models on sentiment classification of parliamentary debate speeches.

To meet these objectives, I compiled ParlVote, a large labelled corpus (34,010 examples) of UK parliamentary debate speeches. These have been labelled at the speech level for use in the evaluation of supervised speech-level sentiment classification systems. In this section, I apply a range of linear and neural machine learning methods and different approaches to text representation to the classification of speeches from this corpus. I also investigate the effects of increases in dataset scale on this task by testing these systems on variously sized subsets of the corpus, and experimented with limiting the length of the input texts to fit the text representation models.

5.2.1 Corpus construction

To create the ParlVote corpus, I downloaded the most recent version of the transcript for each day from May 7th 1997 until November 5th 2019, the last day of the 2017-2019 Parliament.

I developed a tool to retrieve, for each debate, the motion(s) and the utterances of each speaker that voted in the corresponding division. I 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 I compiled manually during development.

I then automatically matched the debates to the corresponding divisions, which are presented in tables by TheyWorkForYou, as described in Section 2.3.1, with votes for ‘aye’ and ‘no’ representing positive and negative sentiment towards a motion, respectively. In order to ensure that the vote labels correspond to the speeches in question, I retained only those debates for which there is exactly one motion and exactly one division. This left 34,010 example speeches in the corpus. For each speech, I included speaker and debate metadata information, as described in Section 2.3.1.

Each example in the final corpus consists of the following fields:

- **debate_id**: identifier (ID) of the debate from which the example is taken.
- **motion_speaker_id**: ID number of the MP who tables the debate motion.
• motion_speaker_name: first name, last name of the MP who tables the motion.

• motion_speaker_party: party affiliation of the MP who tables the motion at the time of the debate.

• motion_text: textual content of the motion.

• speaker_id: ID number of the speaker.

• speaker_name: first name, last name of the speaker.

• speaker_party: party affiliation of the speaker at the time of the debate.

• vote: sentiment polarity class label (1/0) derived from the corresponding division.

• utterance_1 to utterance_n: all utterances made by the speaker in question during the debate.

Pre-processing

To prepare the dataset, I took the following steps. For each utterance, I 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 I judged this 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, I concatenated all the utterances into a single speech.

Unlike the procedure followed in Section 5.2, in this case, I did not remove other parliamentary terms on the intuition that these could in fact be used to indicate speaker sentiment. 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 (they have +0.71 and +0.125 mean sentiment scores, respectively, in sentiment lexicon SentiWordNet\textsuperscript{11}), while in the House of Commons ‘the Honourable Gentleman’ is an obligatory—and therefore neutral—procedural

\textsuperscript{11}SentiWordNet is described in Baccianella et al. (2010).
honorific. I therefore omitted this step, preserving the texts’ original casing. To prepare the text for input to the classification system, I tokenized the speeches and motions.

<table>
<thead>
<tr>
<th></th>
<th>Full corpus</th>
<th>Subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speeches (example units)</td>
<td>34010</td>
<td>33461</td>
</tr>
<tr>
<td>Debates</td>
<td>1995</td>
<td>1995</td>
</tr>
<tr>
<td>Unique speakers</td>
<td>1348</td>
<td>1346</td>
</tr>
<tr>
<td>Parties</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Max. parties per debate</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Min. parties per debate</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean parties per debate</td>
<td>3.63</td>
<td>3.61</td>
</tr>
<tr>
<td>Total tokens</td>
<td>25.74M</td>
<td>26.33M</td>
</tr>
<tr>
<td>Unique tokens</td>
<td>84.89k</td>
<td>81.59k</td>
</tr>
<tr>
<td>Max. utterances per speech</td>
<td>133</td>
<td>—</td>
</tr>
<tr>
<td>Min. utterances per speech</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Mean utterances per speech</td>
<td>3.56</td>
<td>—</td>
</tr>
<tr>
<td>Max. tokens per speech</td>
<td>20967</td>
<td>20730</td>
</tr>
<tr>
<td>Min. tokens per speech</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean tokens per speech</td>
<td>756.76</td>
<td>760.17</td>
</tr>
<tr>
<td>Max. tokens per utterance</td>
<td>7431</td>
<td>—</td>
</tr>
<tr>
<td>Min. tokens per utterance</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Mean tokens per utterance</td>
<td>212.61</td>
<td>—</td>
</tr>
<tr>
<td>Min. speeches per debate</td>
<td>154</td>
<td>149</td>
</tr>
<tr>
<td>Mean speeches per debate</td>
<td>17.05</td>
<td>16.77</td>
</tr>
<tr>
<td>Positive sentiment labels</td>
<td>17993</td>
<td>17721</td>
</tr>
<tr>
<td>Negative sentiment labels</td>
<td>16017</td>
<td>15740</td>
</tr>
<tr>
<td>Government motion examples</td>
<td>18,029</td>
<td>17732</td>
</tr>
<tr>
<td>Opposition motion examples</td>
<td>15981</td>
<td>15729</td>
</tr>
</tbody>
</table>

Table 5.8: 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 I use for sentiment classification experiments.

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 respectively, both versions have fairly balanced sentiment classes. I present the full corpus statistics in Table 5.8. The corpus is available for download at Mendeley Data.

---

5.2.2 Experiments

Approaches

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

Text representations

- **BOW**: I used unigram features as input to the classifiers, with tf-idf feature selection.
- **Word embeddings**: I used pretrained BERT (bidirectional encoder representations from transformers) embeddings (Devlin et al., 2019), which I fine-tuned on the ParlVote data. With the intuition that casing carries important information in parliamentary debates, I used the base, cased model. I 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 further classification layers, as detailed below.

Machine learning classification methods

- **SVM**: commonly used for sentiment analysis in this domain (Balaur et al., 2009; Burfoot, 2008; Burford et al., 2015; Salah, 2014; Thomas et al., 2006; Yessenalina et al., 2010; Yogatama et al., 2015) (and Section 5.1), this is a strong non-neural baseline. I used an SVM with a linear kernel, L2 regularization, and a squared hinge loss function.
- **MLP**: a simple ‘vanilla’ neural network, which has been shown to perform better than SVM in some circumstances on this task (as in Section 5.1). I 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. I 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 classification. In Section 5.1 I 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 MPs from other parties, and classifying them separately. This is attributable 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, and test whether a complex ANN is able to learn these relationships without need for the two-step process described above.

Dataset size:

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, I tested each system configuration on subsets of the corpus of (1) differing sizes, and (2) comprising speeches of either 512 tokens or less, or any number of tokens. I used the following five subsets of the corpus:

- **Full**: the full pre-processed corpus, as described in Section 5.2
- **Medium** (two versions):
  
  * <= 512: all speeches/concatenated speeches + motions of 512 tokens or fewer.
  
  * Any: a random sample of examples (of any length) of the same size as medium (<= 512) (18,253 examples).
- **Small** (two versions):
Any: a random sample of examples (of any length) of the same size as the corpus used in Section 5.1 (1251 examples).

<= 512: as above, but restricted to speeches/speeches+motions of 512 tokens or fewer.

I evaluated systems comprising combinations of these elements 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.

### 5.2.3 Results

Results are presented in Table 5.9. Overall, I found that, with the exception of two classifier/debate model/data subset combinations, the machine learning classification methods outperformed the majority class baselines. Both of these instances occurred 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 given training set, which may simply not be enough for the models to learn from or lead to them overfitting to this limited data.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Debate model</th>
<th>Small (1,251)</th>
<th>Corpus subset</th>
<th>Full (33,461)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Any &lt;=512</td>
<td>Any &lt;=512</td>
<td></td>
</tr>
<tr>
<td>Majority</td>
<td>Motion-ind</td>
<td>40.48</td>
<td>50.71</td>
<td>50.01</td>
</tr>
<tr>
<td>class</td>
<td>Gov/Opp</td>
<td>53.17</td>
<td>49.62</td>
<td>49.09</td>
</tr>
<tr>
<td></td>
<td>Motion+speech</td>
<td>40.48</td>
<td>50.71</td>
<td>50.01</td>
</tr>
<tr>
<td>SVM</td>
<td>Motion-ind</td>
<td>50.00</td>
<td>59.26</td>
<td>61.78</td>
</tr>
<tr>
<td></td>
<td>Govt/opp</td>
<td>51.59</td>
<td><strong>57.94</strong></td>
<td><strong>63.27</strong></td>
</tr>
<tr>
<td></td>
<td>Motion+speech</td>
<td>50.00</td>
<td>60.51</td>
<td>61.82</td>
</tr>
<tr>
<td>MLP</td>
<td>Motion-ind</td>
<td>50.00</td>
<td>59.69</td>
<td>60.05</td>
</tr>
<tr>
<td></td>
<td>Govt/opp</td>
<td>46.83</td>
<td>66.54</td>
<td>65.34</td>
</tr>
<tr>
<td></td>
<td>Motion+speech</td>
<td>44.44</td>
<td>60.84</td>
<td>62.18</td>
</tr>
<tr>
<td>BERT + MLP</td>
<td>Motion-ind</td>
<td>48.41</td>
<td>57.39</td>
<td>60.56</td>
</tr>
<tr>
<td></td>
<td>Govt/opp</td>
<td><strong>64.29</strong></td>
<td><strong>54.38</strong></td>
<td><strong>65.61</strong></td>
</tr>
<tr>
<td></td>
<td>Motion+speech</td>
<td>57.94</td>
<td>61.39</td>
<td><strong>67.31</strong></td>
</tr>
</tbody>
</table>

Table 5.9: Sentiment classification accuracy scores (%) using five subsets of the ParlVote corpus. For each subset, the highest accuracy obtained is highlighted in bold text.
CHAPTER 5. SENTIMENT CLASSIFICATION

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 my findings in Section 5.1, I found no consistent performance gains over the linear classifier from using an ANN.

Text representations Perhaps surprisingly—considering its success on other tasks and domains—I did not see consistent gains from fine-tuning on the BERT embeddings model. Possible explanations for this are that the length of parliamentary speeches make them unsuitable for use with BERT due to it’s low maximum token length, and that linguistic knowledge from the general domain encoded in the BERT model may not transfer well to this domain. However, BERT does obtain the best performance on two of the corpus subsets.

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

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

5.2.4 Discussion and conclusion

While it may be expected that an approach that utilises a neural network and word embeddings approach, such as BERT + MLP would outperform a somewhat simpler approach such as a linear SVM trained on a BOW 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—seconds rather than hours for BOW compared to BERT on the large dataset in the motion-independent setting running on a graphics processing unit (GPU)—it may seem hardly worthwhile to pursue the transformer-based approach.
However, the most prominent setting in which the BERT + MLP model did produce the highest accuracy classification is that of 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 those labels). As I 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 I 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.

5.3 Chapter summary

I have compiled and made available two new corpora of UK parliamentary debate speeches. The first of these includes both manually annotated and automatically sourced sentiment polarity labels. The second includes only the latter type, but is substantially larger. Using these datasets, I tested the effects on sentiment classification performance of a range of combinations of features and language models, machine learning methods, debate models, labelling methods, and corpus subset sizes.

In Section 5.1, I tested hypothesis H1, assessing the validity of class labels derived from speakers’ votes. I concluded that, while there is some difference between these labels and those produced by human annotators, the vote-derived labels have a high degree of validity for use in practice. I also examined hypothesis H2, evaluating the use of a proposed two-step, motion-dependent debate model,
designed to mitigate the polarity shifting effects of motion sentiment when conducting sentiment analysis of debate speeches. I found that this boosted classification performance for all system configurations, and that a Government/opposition motion-labelling method was able to achieve comparable performances.

In Section 5.2, I evaluated sentiment classification systems on a much larger dataset. With regard to hypothesis H4, I found that the use of state-of-the-art transformer-based models did not always lead to the dramatic improvements seen in other domains. This is perhaps due to the tendency for parliamentary speeches to be long, while BERT’s maximum input length is relatively small. I examine this hypothesis further in the experiments in the following chapters.

Results of both sets of experiments strongly suggest that debate modelling should take account of motions to which speeches are addressed. In Chapters 6 and 7, I explore how to determine the topics introduced in the motions, and to detect the speakers’ attitudes towards them.
Chapter 6

Topic identification

‘All issues are political issues, and politics itself is a mass of lies, evasions, folly, hatred and schizophrenia.’

George Orwell, author

Chapter 5 showed that debate motions (tabled proposals) contain important information about the topics and policies discussed in parliamentary debates. In addition to acting as polarity shifters, which can effect the ability of classifiers to correctly identify features indicative of positive or negative sentiment and stance, they are the focus of the opinions expressed during subsequent speeches. They are therefore crucial for any understanding and processing of the content of MPs’ spoken utterances, and analysis of the opinions and positions they convey.

In this chapter, I therefore present work on automatic identification of the topics and policies discussed by debate participants in debate motions. Initially, in Section 6.1 I outline why existing generic topic labels and commonly used approaches to discovering the latent topics in texts are unsuitable for the aims of this project.

I examine research hypothesis H3: *Supervised machine learning classifiers can predict the manually applied opinion-topic labels of debates from textual (and metadata) features of the debate motions and speeches in a multiclass setting.* I present two sets of experiments conducted on two different datasets. Firstly, in Section 6.2 I evaluate the use of multi-label, multi-task classification to automatically label motions with one of a set of pre-existing crowd-sourced policy labels,
and discuss the strengths and limitations of this approach. Secondly, in Section 6.3, I create a new corpus of debate motions labelled with policy preference labels. This is based on a framework from political science concerned with the annotation of party political manifestos. On this dataset, I compare the performance of a range of heuristic methods and state-of-the-art supervised classification approaches to the automatic labelling of the motions with these categories.

6.1 Topics in parliamentary debates

In this section, I examine the utility of a pre-existing schema for labelling debates with topics, and an unsupervised approach to topic extraction.

6.1.1 Pre-existing ontology of topics

There exists a pre-existing systematic ontology of topic labels for parliamentary materials, produced by the House of Commons Library for the categorisation of their research briefings. However, the topics it consists of, such as Education or Transport, are general and neutral, and therefore are not the type of topics that are likely to be the targets of speakers’ opinions. That is, MPs do not tend to take a position for or against such generic topic areas. Rather, they express support or opposition towards particular policies associated with different topics. What I therefore aim to extract from the motions in this project are opinion-topics that provide information about these policies—the speakers’ policy preferences.

6.1.2 Topic modelling

A commonly used approach to the extraction of topic information from text documents is that of topic modelling, about which a large body of research exists, much of which concerns variations of the LDA algorithm of Blei et al. (2003). However, initial experiments running topic modelling on the Hansard transcripts revealed this approach to be unsuitable for the purposes of this project for similar reasons as the House of Commons Library topic labels above. I ran LDA on a subsection of the Hansard corpus, example output of which can be seen in Table 6.1.

---

1 A full list of House of Commons Library topics is available at [https://www.parliament.uk/topics/topical-issues.htm](https://www.parliament.uk/topics/topical-issues.htm), accessed 16 July 2020.
CHAPTER 6. TOPIC IDENTIFICATION

Top five word stems in topic | Interpreted topic label
--- | ---
1. polic, offic, minist, crime, forc | Law & order
2. care, school, secretari, social, fund | Welfare & education
3. cost, invest, govern, time, world | Government spending
4. hous, bill, govern, time, member | Parliament – procedures
5. rural, govern, busi, european, economi | Economy
6. hon, friend, right, gentleman, agre | Parliament – people
7. peopl, vote, minist, countri, nation | Democracy
8. work, peopl, prison, servic, need | Prison service
9. women, peopl, pension, work, man | Pensions
10. council, govern, tax, author, increas | Local government funding

Table 6.1: Output of structured topic modelling on the Hansard debates with \( n = 10 \) topics.

While some thematic areas that are discussed in debates emerge, and it is possible to interpret these and label them (performed here manually), the resulting topics, such as Law & order or Economy, again appear to be both general and neutral. While a topic modelling approach has been used with Hansard data, (as in Nanni et al. (2019), for instance), it is unsuitable for the purposes of this project, in which I seek to extract subjective opinion-topics. Alternative forms of topic labelling are therefore required.

To this end, I have made use of two sources of data. In Section 6.2, I investigated the use of crowd-sourced labels produced by the Public Whip website, which refers to the opinion-topics as policies. In Section 6.3, I applied the Manifesto Project (MARPOR) policy preference labelling scheme to this domain.

The experiments in this chapter are therefore focused on the following two objectives:

1. Investigation of the use of pre-existing crowd-sourced ‘policy’ labels to train a multi-class, multi-label classifier to identify polarised opinion-topics in parliamentary debate motions (Section 6.2).

2. Investigation of the adaptation to the parliamentary debate domain of a pre-existing coding scheme for the labelling of party political manifestoes for the automatic identification of policy preferences in debate motions (Section 6.3).
6.2 Multi-class, multi-label classification of motion policies

Debate motions are so hard for ordinary citizens to make sense of that parliamentary monitoring organisations like Public Whip and TheyWorkforYou produce annotated versions and manually written summaries of them. These are the contributions of crowd-sourced volunteers with domain expertise or interest. Example 6.1 shows a Public Whip entry in which a motion has been summarised and annotated with a policy label. Note that while this particular example motion is short and relatively straightforward to understand, many motions are comprised of several clauses and can refer to many external entities with which most readers may not be familiar.

Motion: That this House endorses the Government’s formal application to rejoin 35 European Union Justice and Home Affairs measures, including the European Arrest Warrant.

Summary: In July 2013 the UK took a decision to opt out of 130 European Union measures in the field of police co-operation and judicial co-operation in criminal matters. The UK Government identified 35 measures it wished to rejoin prior to the opt out decision taking effect on 1 December 2014. These 35 European Union measures, to which the motion relates, are listed, and explained, in the Explanatory memorandum to The Criminal Justice and Data Protection (Protocol No. 36) Regulations 2014.

Policy: #1065: “European Union Integration — For”

In this Section, I describe experiments, in which I use the policy annotations as opinion-topic labels. I also exploit the fact that these topic labels inherently incorporate information regarding the proposers’ opinions towards those policies—that is, their policy positions: whether they support or oppose them. I investigate whether, by correctly identifying a motion’s policy category, we can also determine its position towards the policy in question, in effect obtaining

opinion analysis of the motions ‘for free’. I compare the output of this approach to human produced opinion polarity labels. To this end, I created a dataset of UK parliamentary debate motions labelled with both topic and opinion polarity.

**Research objective** The research objective of this section is therefore the evaluation of multiclass, multilabel classification using the PublicWhip’s *policy*, labels as opinion-topic classes.

A version of the work presented in this section have previously been published in [Abercrombie and Batista-Navarro (2018c)](https://www.publicwhip.org.uk/policies.php).

### 6.2.1 Opinion-topic labelling

The PublicWhip website maintains a list of debates (titles plus vote outcomes) organised under ‘*policies*’ (see definition in Section 6.1). Examples include *European Union – For* and *Stop climate change*. Under each *policy*, members of the public are invited to submit debates—in the form of motions with vote outcomes—which match these descriptions.

I made use of these categorisations as labels for supervised topic classification. In many, if not most, cases, it would not be straightforward to determine a motion’s *policy* label from the debate title—for example, under the PublicWhip *policy* ‘More Powers for Local Councils’, debate titles include ‘High Streets’, ‘Housing’, ‘Fixed Odds Betting Terminals’, and ‘Local Bus Services’, which do not necessarily indicate a connection with the policy area of devolved power. Similarly, the text of a motion alone does not necessarily reveal its topic, with many motions consisting purely of procedural or legislative language, such as ‘*That the Bill be read a Second time*’. As a result, a human reader would often require access to the title, motion, and sometimes other information found elsewhere in a debate—or even from an external source—in order to determine the motion’s opinion-topic.

In the PublicWhip categorisation framework, the *policy* labels represent both an opinion-topic and a polarised position towards it. However, the latter is a reflection of the vote outcome of the debate, not necessarily the position expressed in the motion. Under this labelling scheme, if a motion proposed in support of a

---

policy position is rejected, it will be labelled with a policy that reflects opposition to that position. For instance, in Example 6.2, the motion, from a debate entitled ‘Opt Out from European Union Police and Criminal Justice Measures’, has the policy label #1065: “European Union Integration — For. In fact, the motion is firmly against EU integration, but the fact that it was rejected by the House, explains why it has been given this label.

Motion: That this House believes that the UK should opt out of all EU police and criminal justice measures adopted before December 2009 and seek to rejoin measures where it is in the national interest to do so and invites the European Scrutiny Committee, the Home Affairs Select Committee and the Justice Select Committee to submit relevant reports before the end of October, before the Government opens formal discussions with the Commission, Council and other Member States, prior to the Government’s formal application to rejoin measures in accordance with Article 10(5) of Protocol 36 to the TFEU.

Policy: #1065: “European Union Integration — For”

Policy vote: Minority

In a further layer of complexity, the Public Whip also provides motions with a ‘policy vote’ annotation. This represents the contributor’s assessment of how somebody who supports the relevant policy ‘would have voted’. This field is completed with one of three additional class labels, indicating whether a vote with the majority, or with the minority (or to abstain) would support the policy in question. All in all, this means that each labelled example has two potential polarity shifters—the vote outcome and the policy vote—which need to be taken into account if these policies are to be used as labels for opinion polarity analysis.

In order to interpret the labels, I treat the ‘majority’ motions as being labelled according to the vote outcome—those which were ‘approved’ by the vote are positive, while those which were ‘rejected’ are negative—and the ‘minority’ labels as polarity shifters—that is, the tag ‘minority’ reverses the label derived from the outcome, while a ‘majority’ tag preserves it. Figure 6.1 demonstrates this process.
CHAPTER 6. TOPIC IDENTIFICATION

135

Motion policy

Approved

Rejected

'majority'

'minority'

'majority'

'minority'

positive

negative

negative

positive

Outcome Policy Vote Opinion

Figure 6.1: Interpretation of policy labels for opinion analysis. In the PublicWhip data, for each of its policy labels, a motion has two further tags—outcome and policy vote—that can potentially reverse its opinion polarity.

6.2.2 Data

I constructed a dataset of UK parliamentary debate motions proposed in the House of Commons between 1997 and 2018. I matched these with the corresponding policies from the Public Whip for use as labels for supervised opinion-topic classification. I therefore included only those motions which have been classified by policy on the website. In order to provide sufficient examples to train a classifier I used only those debates for which there exist at least 20 examples per policy label.

Because, for example, a debate may have been categorised with both the specific policy ‘Higher taxes on alcoholic drinks’, and the more general label ‘Increase VAT’ (Value Added Tax), motions can be included in more than one policy category and therefore have multiple class labels. The final dataset includes 13 different policy topic labels, with each applied to a minimum of 24 and a maximum of 129 motions ($\mu = 46.6$) (see Table 6.2). 14 of the motions have two labels, while the remaining 578 have just one.

4The debate transcripts were obtained from https://www.theyworkforyou.com/pwdata/scrapedxml/debates, accessed 16 July 2020.
Table 6.2: The number of motions labelled with each policy class label in the final dataset.

In addition to the crowdsourced labels, I provide a second set of manually annotated opinion-polarity labels. For these, I read each example (motion, title, and supplementary information), and applied either positive or negative labels according to the opinion I perceived was expressed towards the policy in question.

I observed that the motions in this dataset broadly follow one of the following three formats:

1. ‘That this House {verb} {argument}; {verb} {argument}; ... and {verb} {argument}’
   –where the motion may contain several clauses.

2. That the {legislation} be now read {a Second/the Third} time.
   –where {legislation} is a Bill, amendment etc.

3. ...amendment {number}, page {number}, line {number}, leave out {phrase} and insert ‘{phrase}’, in which an attempt is made to alter the wording of a previously tabled motion.

Motions of types 2 and 3 contain very little topic information, in which cases it seemed likely that it would be necessary to make use of cues in the debate title or additional information provided in the transcript in order to determine the correct policy. As further features for machine classification, I therefore included
the following metadata information from the transcripts as well as the textual content of the motions:

- **motion_speaker_name**: Some MPs are more or less likely to speak on various topics, depending on their interests and position.
- **motion_party**: Party affiliation of speakers is likely to be an indicator of both interest in topics and policy positions.
- **debate_title**: Titles are often, but not always, related to policy vote topics.
- **additional_information**: Information such as the names of relevant documents or explanations of amendments is often included in the transcripts, preceding the motion.

The final dataset consists of 592 labelled debate motions.

### 6.2.3 Experiments

Data pre-processing consisted of removal of stopwords, lowercasing and stemming of textual data, and binarization of metadata information. I provided each example in the dataset with the 13 class labels using one-hot encoding (whereby each label is a 13-element binary array).

In order to detect the topics of debate motions I employed a supervised machine classification approach. For this, I investigated the use of all possible combinations of the following features:

- **Textual features**: n-grams (uni-, bi-, and trigrams) from the debate titles, motions, and supplementary information provided by the crowdsourced annotators.
- **Metadata features**: speaker names and party affiliations.

As some motions have more than one topic label, I applied one-vs-the-rest classification on a randomised 90%—10% train-test split of the data. Classification was performed using a multilabel implementation of an SVM.

---

5 The Motion Policies Corpus is available at [https://data.mendeley.com/datasets/j83yzp7ynz/1](https://data.mendeley.com/datasets/j83yzp7ynz/1), accessed 16 July 2020.
6.2.4 Results

Results are presented in Table 6.2. Because there were 13 different classes, and therefore highly imbalanced datasets for each round of one-vs-the-rest classification, I used the (macro-weighted) F1 score as a performance metric. Strongest performance was achieved using \( n \)-gram features from both the debate motions and titles (F1 = 77.0).

![Figure 6.2: F1 score for different combinations of features used for topic classification. Highest performances (lighter, orange bars) are achieved when textual features derived from the debate titles are included.](image)

Overall, use of the debate titles, with or without metadata features, produced the highest F1 scores, while the addition of other textual features did not generally lead to improvement, and in some cases resulted in performance losses.

The text content of the motions themselves do not appear to provide particularly useful features for topic detection. Many consist solely of procedural terms that give no indication of the topics under discussion—such as motion types 2 and 3 (described in Subsection 6.2.2). Indeed, only 121 (20.8 per cent) motions are of the more easily interpreted type 1.

Of the metadata information used, speaker name is a more indicative feature of topic than party affiliation. This reflects the fact that each party is represented in most policy categories, but that individual MPs tend to be strongly associated with just a few, or in most cases, one single topic related to their particular role—of
234 MPs in the dataset, 163 (69.7 per cent) propose motions on only one policy, and only one is represented in more than four.

As the original threshold of 20 for the minimum number of examples per policy in the dataset was somewhat arbitrary, I also tested the system with a range of different thresholds (see Figure 6.3). As this threshold decreases and the number of different topic classes increases, the F1 score drops, indicating that it may be challenging to obtain good results with a larger corpus and a greater number of topics.

![Figure 6.3: F1 scores for classification with different thresholds (θ) for the minimum number of example motions. As this threshold is lowered, the number of topic classes (n) increases and performance decreases.](image)

6.2.5 Discussion

Considering the small number of training examples for each class, reasonable results are obtained using these labels for topic classification. However, it should be noted that many of the policy classes in this dataset feature debates with similar or even identical titles, in which cases the classifier is trained and tested on very similar data. While this is a common scenario in Parliament—the same pieces of legislation are often debated multiple times and often revisited year after year—it is unlikely that this approach would perform well on new, completely unseen examples from future debates, in which completely different topics and policies may be discussed.
Although the opinion-topic classification results are promising, it should be noted that the system used to interpret motion opinion from policies relies on the use of additional, manually applied policy vote tags. For use with future, unseen examples that do not have such tags, it would be necessary to reorganise the way that PublicWhip’s policies are created, splitting those labelled ‘majority’ and ‘minority’ into different for and against Policy categories. While this may complicate topic identification (for each current policy, there would now be two similar ones—a negative and a positive), it might also make metadata features more predictive as different MPs/parties tend to hold different policy positions.

\section*{6.2.6 Conclusion}

Opinion-topic labels derived from the PublicWhip policies can be used to train a classifier to achieve good performance in classifying the policy topics of parliamentary debate motions. These categories, which incorporate inbuilt policy position information, also appear to be reliable markers of opinion polarity, suggesting that we can use these labels to simultaneously obtain opinion analysis ‘for free’.

However, despite these promising results, this approach suffers from a number of weaknesses: (1) the number of available labelled motions is limited, with most policies having insufficient examples with which train a classifier; (2) it is likely to be unable to deal with new examples from future, unseen debates; (3) if the Public Whip policies were to be used in this way at a larger scale, they would have to be reorganised to omit both abstentions and those labelled ‘rejected’ or ‘minority’, which would require more systematic labelling from the crowdsourced annotators; and (4) while a speaker’s party affiliation might be expected to be a strong indicator of their policy position, the way the Public Whip’s policies are organised by vote outcome means that this feature is not a strong indicator of topic using these labels, and exploitation of this this would require further restructuring of the labelling system.

In the next section, I therefore explore an alternative framework for labelling opinion-topics in debate motions.
6.3 Policy preference detection in debate motions

In this section, I present the results of experiments in which I attempt to apply to debate motions the coding schema devised by MARPOR. This schema is convenient because, unlike the framework used in Section 6.2: (a) it is systematic, having been developed by political scientists over a 40 year period, (b) it is comprehensive and designed to cover any policy preference that may be expressed by any political party in the world, (c) it has been devised to cover any policies that may arise in the future, and (d) there exist many expert-coded examples of manifestos, which we can use as reference documents and/or for validation purposes.

Research objective The objective of this section is to evaluate the use of a coding framework and data produced by experts from the field of political science for the labelling and automatic identification of the policy preferences (policy-focused opinion-topics) expressed in debate motions.

Parts of this section have been published in Abercrombie et al. (2019).

6.3.1 Background

The concept of policy preferences is widely used in political science literature (e.g. Budge et al. 2001; Lowe et al. 2011; Volkens et al. 2013) to represent the positions of political actors expressed in text or speech. The Manifesto Project (MARPOR) is an ongoing venture that spans four decades of work in this area and consists of a collection of party political documents annotated by trained experts with codes (labels) representing such preferences. Organised under seven ‘domains’, the coding scheme comprises 57 policy preference codes, all but two of which (408: Economic goals and 000: No meaningful category applies) are ‘positional’, encoding a positive or negative position towards a policy issue (Mikhaylov et al. 2008). Indeed, many of these codes exist in contrastive, polar opposite pairs, such as 504: Welfare State Expansion and 505: Welfare State Limitation. The manifestos are labelled labelled at the quasi-sentence level—that is, units of text that span a sentence or part of a sentence, and which have been judged by the annotators to contain ‘exactly one statement or “message”’
In Example 6.3, a single sentence from a manifesto has been annotated as four quasi-sentences (QSs), each of which has been given a different policy preference label:

QS1: To secure your first job we will create 3 million new apprenticeships;

411: Technology and Infrastructure

QS2: take everyone earning less than £12,500 out of Income Tax altogether

404: Economic Planning

QS3: and pass a law to ensure we have a Tax-Free Minimum Wage in this country;

412: Controlled Economy

QS4: and continue to create a fairer welfare system where benefits are capped to the level that makes work pay — so you are rewarded for working hard and doing the right thing.

505: Welfare State Limitation

6.3.2 Data

To construct the corpus, I made use of the data sources described below.

The Manifesto Project

I used annotated manifestos (a) as reference texts for labelling of debate motions by similarity matching, and (b) to train a neural network for cross-domain classification of the motions. I downloaded all fifteen of the annotated UK (including Northern Ireland)
manifestos from the Manifesto Corpus Version 2018-1—those that have been coded under version 4 of the coding scheme. In this subset, the number of UK manifesto quasi-sentences labelled with codes in each domain varies considerably (see Table 6.3).

---

7An excerpt from the Conservative Party manifesto 2015.

8While part of the UK, distinct political parties operate in Northern Ireland, and MARPOR treats it as a separate territory.

CHAPTER 6. TOPIC IDENTIFICATION

Table 6.3: The numbers of policy preference codes, quasi-sentences (QSs) coded under each domain in the UK manifestos that I use as reference texts and training data, and debate motions and quasi-sentences that I label under each domain in the motion policy preference corpus.

These manifestos were published by a variety of political parties for election campaigns over an 18 year period (see Table 6.4). The most prevalent code in these manifestos is 504: Welfare State Expansion (2691 examples), and the least used is 103: Anti-Imperialism (3 examples). Two codes, 102: Foreign Special Relationships: Negative and 415: Marxist Analysis: Positive, do not appear at all in manifestos from the United Kingdom.

Table 6.4: The parties and years of publication of the manifestos that I use as reference texts and training data, and the number of labelled quasi-sentences (QSs) by party in this subset of the manifesto data.

I downloaded all files from May 7th 1997 (the start of that year’s session of Parliament) to February 28th 2019 from TheyWorkForYou. From these I extracted 1,156 motions together with the titles of the debates and the dates on
which they were tabled. I manually removed procedural motions (those concerned solely with the workings of Parliament) from the dataset as these do not concern policy preferences and have no equivalents in political manifestos.

In order to approximate the format of the MARPOR data, and to investigate policy preference detection at different levels of granularity, I divided each motion into smaller units. For convenience, I approximated quasi-sentences in the Hansard data by automatically dividing motions into clauses, which are always separated by semi-colons in the transcripts.

6.3.3 Annotation

I adapted the MARPOR manifesto Coding Instructions (Werner et al., 2011) to provide guidelines for the annotation of debate motions.\textsuperscript{10} I used version 4 of these instructions because, although a more recent, more finely grained version exists, there are as yet few example manifestos coded under the newer scheme. To complete the annotation task, I recruited three political science Master’s students from the University of Mannheim, who worked for a total of 40 hours each over a two month period.

Debate motions

Annotations were carried out in two stages: an initial training phase, followed by labelling of the main dataset. I used the coding instructions of version 4 of the Manifesto Project handbook\textsuperscript{11} supplemented by parliamentary debate motion-specific guidelines including notes based on the annotators’ discussions during training. I make these adapted guidelines available in Appendix B.

For the training phase, after being introduced to the data and the coding instructions, the annotators individually labelled three batches of motions and their quasi-sentences. In addition to labelling each of these with one of the codes, they were instructed to note examples which they found difficult to decide upon. Between each batch we met to discuss these instances, as well as other examples on which the annotators disagreed, adding notes to the annotation guidelines based on the observations made. Inter-annotator agreement during training ranged

\textsuperscript{10} Annotation guidelines available in Appendix B
from ‘fair’ to ‘substantial’, following common interpretation of Fleiss’ kappa scores (Landis and Koch, 1977) (see Table 6.5).

<table>
<thead>
<tr>
<th>Iteration</th>
<th>No.</th>
<th>k</th>
<th>Agreement</th>
<th>No.</th>
<th>k</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training 1</td>
<td>15</td>
<td>0.41 ‘moderate’</td>
<td>60</td>
<td>0.35 ‘fair’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training 2</td>
<td>12</td>
<td>0.65 ‘substantial’</td>
<td>60</td>
<td>0.56 ‘moderate’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training 3</td>
<td>16</td>
<td>0.48 ‘moderate’</td>
<td>60</td>
<td>0.40 ‘fair’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.5: Common interpretation of annotator agreement (Fleiss’s kappa) at two levels of granularity during three iterations of training and development of annotation guidelines for labelling debate motions with the MARPOR codes.

The final corpus includes manual annotations of 386 motions comprised of 1,683 quasi-sentences. The majority of these have been labelled by two of the three annotators. Inter-annotator agreement is within the ranges generally interpreted as being ‘moderate’ to ‘substantial’ (see Table 6.6). The slightly higher agreement at the quasi-sentence level than on overall motion labels suggests that it may be difficult in some cases to select a single policy preference code for a whole motion. A subsection of the corpus (41 motions, 180 quasi-sentences) was labelled by all three annotators. Fleiss’ kappa scores for this subsection are 0.46 at both levels, which indicates ‘moderate’ agreement. Following Pustejovsky and Stubbs (2012), I applied the gold standard label for each example, adjudicating on any examples that were the subject of inter-annotator disagreement.

**Manifestos**

To validate the labelling procedure, and for comparison with other work, I also asked the annotators to label a small quantity (120) of quasi-sentences from the Manifesto Project. I calculated Fleiss’ kappa for these annotations to be 0.48, which is comparable to that obtained on the main dataset of debate motions, and higher than those reported by Mikhaylov et al. (2008) on annotation of manifests.

Again, the annotators were asked to mark any examples which they considered to be difficult to decide upon. Agreement (Fleiss’ kappa) on these ‘difficult’ cases is only 0.17, with only one example marked as such by all three annotators. In this figure represents examples with ‘gold standard’ labels. The corpus also includes examples labelled by a sole annotator (‘silver standard’) and further unlabelled motions (see Table 6.7).
CHAPTER 6.  TOPIC IDENTIFICATION

Annotators | No. of examples | $k$
---|---|---
Motion | All 3 | 41 | 0.46
QS | All 3 | 180 | 0.46
Motion | 1 & 2 | 139 | 0.51
| 2 & 3 | 155 | 0.50
| 1 & 3 | 169 | 0.49
| All pairs | 463 | 0.50
Motion | 1 & 2 | 622 | 0.58
| 2 & 3 | 650 | 0.51
| 1 & 3 | 731 | 0.62
| All pairs | 2003 | 0.58

Table 6.6: Fleiss’ kappa scores for three-way agreement and Cohen’s kappa scores for two-way agreement on the debate motions dataset.

In this case, two of them used the ‘correct’ MARPOR gold label, while the third annotator applied a different code from the same domain. Of the 47 examples (39.2%) on which all three annotators agree, 36 of these agree with the gold label (30% of the total). Domain-level agreement is 0.56, which is also similar to that achieved on the debate motions. Overall, annotation of manifestos and debate motions appears to involve a similar degree of difficulty, with much of the disagreement due to individual annotators’ interpretation of the texts, rather than any clear pattern of ambiguous examples.

The Motion Policy Preference Corpus

The corpus is available for download at https://madata.bib.uni-mannheim.de/308. The number of labelled and unlabelled examples it contains can be seen in Table 6.7. For the gold-labelled data, motions range in length from one to 13 quasi-sentences (mean = 4.3), with each of these consisting of between four and 163 tokens (mean = 28.7).

For the gold standard-labelled data, the number of motions labelled with codes from each domain is presented in Table 6.8.

---

Table 6.7: Number of labelled and unlabelled examples in the Motion Policy Preference Corpus.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Labelled Examples</th>
<th>Unlabelled Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARPOR</td>
<td>47</td>
<td>456</td>
</tr>
<tr>
<td>COREMAP</td>
<td>47</td>
<td>456</td>
</tr>
<tr>
<td>DIPNOTES</td>
<td>47</td>
<td>456</td>
</tr>
</tbody>
</table>

---

13 URL accessed 31 August 2020.
CHAPTER 6.  TOPIC IDENTIFICATION

<table>
<thead>
<tr>
<th>Labels</th>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Motion</td>
</tr>
<tr>
<td>Gold standard</td>
<td>386</td>
</tr>
<tr>
<td>Silver standard</td>
<td>87</td>
</tr>
<tr>
<td>Total labelled</td>
<td>473</td>
</tr>
<tr>
<td>Unlabelled</td>
<td>593</td>
</tr>
<tr>
<td>Overall total</td>
<td>1,066</td>
</tr>
</tbody>
</table>

Table 6.7: Statistics for the motion policy preference corpus. Gold standard examples have been labelled by two or three annotators initially and adjudicated on in a final round of annotation. Silver standard examples have been labelled by a single annotator only.

<table>
<thead>
<tr>
<th>Code</th>
<th>Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Motion</td>
</tr>
<tr>
<td>1: External Relations</td>
<td>50</td>
</tr>
<tr>
<td>2: Freedom &amp; Democracy</td>
<td>30</td>
</tr>
<tr>
<td>3: Political System</td>
<td>47</td>
</tr>
<tr>
<td>4: Economy</td>
<td>87</td>
</tr>
<tr>
<td>5: Welfare &amp; Quality of Life</td>
<td>118</td>
</tr>
<tr>
<td>6: Fabric of Society</td>
<td>33</td>
</tr>
<tr>
<td>7: Social Groups</td>
<td>21</td>
</tr>
<tr>
<td>0: No meaningful category</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.8: Domains of codes assigned to debate motions in the gold-labelled dataset.

### 6.3.4 Experiments

I investigated two ways of automatically labelling debate motions with the codes from the Manifesto Project: (1) similarity matching, which provides an unsupervised baseline; and (2) supervised classification. I tested these both at the quasi-sentence level and additionally experimented with similarity matching methods at the whole motion level, where the lack of sufficient training data prevents application of supervised learning methods. In pre-processing I filtered out any motions that have gold standard labels that appear less than ten times in the corpus, leaving 370 motions and 1,634 quasi-sentences, each annotated with one of the 32 remaining class labels.
CHAPTER 6. TOPIC IDENTIFICATION

Similarity matching

I tested two methods of matching debate motions to codes from the Manifesto Project, comparing a baseline of unigram overlap scores with cosine similarity measurement. In each case, I measured the similarity of the list of tokens \( A = A_1, A_2, ..., A_n \) in each motion or quasi-sentence text and the list of tokens in each collection of concatenated manifesto extracts \( B = B_1, B_2, ..., B_n \).

For unigram overlap, a simple count was performed of the union of the sets of tokens from \( A \) and \( B \). For the latter method, each document was converted to a [Tfidf] representation, and cosine similarity between the motion and manifesto documents was calculated as:

\[
\frac{A \cdot B}{||A|| ||B||}
\]

With both of these approaches, I explored the use of the following combinations of sources of textual unigram features: the debate titles, which I had previously found to be highly predictive of a motion’s opinion-topic in a supervised classification setting (see Section 6.2), the debate motions themselves, and both the titles and motions together.

Supervised Classification

I tested a range of supervised machine learning algorithms for the policy preference classification task, ranging from traditional approaches to recently developed pre-trained deep language representation models. I was particularly interested in assessing the performance of such approaches: (1) despite the limited training data available (around 1,600 motion quasi-sentences); and (2) in a cross-domain application (training on over 16,000 manifesto quasi-sentences, and testing on the motion quasi-sentences).

First, I examined the performance of an SVM trained using lexical [Tfidf] or word embedding (w-emb) features, which act as strong traditional baselines. I tested both pre-trained general purpose word embeddings from fastText\(^{15}\) (Mikolov et al., 2018) and in-domain vectors generated on the Hansard transcripts from Nanni et al. (2019).

I also report the results of a widely adopted neural network baseline for topic classification (see, for instance, Glavaš et al. (2017a) and Subramanian et al. (2018) in the context of manifesto quasi-sentences classification): a CNN with a

\(^{14}\)Preliminary experiments showed unigram overlap to perform favourably compared with Jaccard similarity.

\(^{15}\)fastText website: [https://fasttext.cc](https://fasttext.cc) accessed 27 July 2020.
single convolutional layer and a single max-pooling layer. I again tested the CNN with general purpose and in-domain embeddings.

As final skyline comparisons, I present the performance of (1) a pre-trained BERT \textit{(large, cased)} model \cite{devlin2019bert}, with a final classification layer; and (2) the same pre-trained BERT model, with a CNN. I additionally experimented with the latter two models in a fine-tuning setting: after training on manifestos, they have been further fine-tuned on motions.

I tested all approaches with a 80/20 split of the dataset, and trained all the neural models for three iterations, as recommended in \cite{devlin2019bert}.

### 6.3.5 Results

I evaluated the predicted labels of each experimental model against the gold standard labels produced by the annotation process. For the machine learning methods, I report F1 scores with both macro and micro weightings in order to offer an understanding of the quality overall, as well as for the different classes.

**Motions: similarity matching**

I evaluate labelling of motions by similarity matching at two levels of granularity: quasi-sentence and whole motion. Cosine similarity matching comfortably outperforms the baseline at both levels of granularity and at both MARPOR’s \textit{policy} and \textit{domain} levels (see Figure 6.4).

Unlike the findings of Section 6.2 using the Public Whip \textit{policies}, in most settings, I did not find the debate titles to be as powerful indicators of class labels as features derived from the texts of the motions, perhaps due to the larger set of class labels containing more similar (from the same MARPOR domain) policy preference codes.

Best performances at both policy and domain levels (F1 macro = 0.59) are obtained using \textit{tf-idf} features derived from both motion titles and texts, although performance using the texts only is comparable. For most combinations of feature input and similarity measurement method, F1 scores are around twice as good at the domain level as at the policy level.
Figure 6.4: F1 macro scores for unigram overlap and cosine similarity matching at MARPOR’s policy and domain levels using textual features from whole motions. Use of cosine similarity leads to markedly better performance than unigram overlap, and the best performance is achieved using features derived from both the titles and motion texts at policy and domain levels.

Motions: classification of quasi-sentences

I tested the supervised systems at the quasi-sentence level and at the two levels of class label granularity (policy and domain), which allows for comparison of the results with previous work on MARPOR data (for example, Zirn et al. (2016)). As can be seen in Table 6.10, the use of machine learning methods generally (but not always) leads to a substantial improvement (especially for Micro F1), in comparison to the heuristics that we have discussed above.

Concerning the SVM and CNN baselines, training the classifiers on the large collection of annotated manifestos and then applying them to the motions does not lead to improvements in comparison to the performance of the same architectures on the motions alone. Similarly, it is noticable that in most cases the use of in-domain embeddings does not improve the results. These two findings might be due to the fact that the style of communication and vocabulary of manifestos and debate speeches are very different. The size of the training data may also play
a role, as can be noticed in particular with the weak performances of the CNNs, especially in comparison to more traditional approaches; in the next section, I return to this issue.

Finally, to further confirm the potential of BERT even in tasks which involve many labels, a lack of training data, and a very specific style of communication, use of this state-of-the-art architecture produces a clear improvement over all other systems when fine-tuned on the debate motions.

### Manifestos: classification of quasi-sentences

<table>
<thead>
<tr>
<th>Model</th>
<th>Text representation</th>
<th>Policy</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Macro</td>
<td>Micro</td>
</tr>
<tr>
<td>SVM</td>
<td>Tf-idf</td>
<td>0.39</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Domain w-emb</td>
<td>0.35</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Wiki w-emb</td>
<td>0.38</td>
<td>0.54</td>
</tr>
<tr>
<td>CNN</td>
<td>Domain w-emb</td>
<td>0.28</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Wiki w-emb</td>
<td>0.27</td>
<td>0.44</td>
</tr>
<tr>
<td>BERT</td>
<td>Large, cased</td>
<td>0.42</td>
<td>0.58</td>
</tr>
<tr>
<td>BERT + CNN</td>
<td>Large, cased</td>
<td>0.42</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 6.9: F1 scores for classification of party political manifestos at the quasi-sentence level.

As a final comparison of the presented systems for quasi-sentence classification, I report their performance on the corpus of 16 thousand manifesto quasi-sentences, again with an 80/20 train-test split. The results (see Table 6.9) are consistent with the performance of supervised pipelines on the Manifesto Corpus presented in previous literature ([Glavaš et al., 2017a](#), [Subramanian et al., 2018](#), [Zirn et al., 2016](#)) and in line with the performances obtained on the motion corpus shown in Table 6.10.

Interestingly, it can be noticed that the CNNs performed relatively poorly on the collection, even with ten times as much training data. This could be due to a necessity to extend the architecture (for example, by adding more convolutional layers) rather than a simple lack of training data. Conversely, traditional SVM baselines offer reasonable results, and the best performances were obtained when employing BERT.
### Table 6.10: F1 scores for similarity matching and classification of debate motions at the quasi-sentence level. Word embeddings used were trained on Hansard (Dom. emb) or Wikipedia (Wiki emb).

<table>
<thead>
<tr>
<th>Model</th>
<th>Text repr.</th>
<th>Data src.</th>
<th>Policy</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Macro Micro</td>
<td>Macro Micro</td>
</tr>
<tr>
<td>Unigram overlap</td>
<td>BOW</td>
<td>Motion titles</td>
<td>0.09 0.15</td>
<td>0.26 0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motions</td>
<td>0.09 0.21</td>
<td>0.23 0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>0.10 0.23</td>
<td>0.25 0.39</td>
</tr>
<tr>
<td>Cosine similarity</td>
<td>Tf-idf</td>
<td>Motion titles</td>
<td>0.23 0.34</td>
<td>0.44 0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Motions</td>
<td>0.30 0.36</td>
<td>0.50 0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>0.32 0.41</td>
<td>0.51 0.56</td>
</tr>
<tr>
<td>SVM</td>
<td>Tf-idf</td>
<td>Motions</td>
<td>0.33 0.48</td>
<td>0.58 0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.29 0.40</td>
<td>0.53 0.56</td>
</tr>
<tr>
<td></td>
<td>Dom. emb</td>
<td>Motions</td>
<td>0.32 0.50</td>
<td>0.53 0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.25 0.41</td>
<td>0.45 0.53</td>
</tr>
<tr>
<td></td>
<td>Wiki emb</td>
<td>Motions</td>
<td>0.35 0.51</td>
<td>0.55 0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.21 0.38</td>
<td>0.45 0.52</td>
</tr>
<tr>
<td>CNN</td>
<td>Dom. emb</td>
<td>Motions</td>
<td>0.15 0.38</td>
<td>0.58 0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.19 0.30</td>
<td>0.37 0.51</td>
</tr>
<tr>
<td></td>
<td>Wiki emb</td>
<td>Motions</td>
<td>0.13 0.29</td>
<td>0.50 0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.21 0.36</td>
<td>0.48 0.56</td>
</tr>
<tr>
<td>BERT</td>
<td>Lg, cased</td>
<td>Motions</td>
<td>0.26 0.47</td>
<td>0.42 0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.32 0.47</td>
<td>0.52 0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ fine-tune</td>
<td>0.39 0.50</td>
<td>0.60 0.67</td>
</tr>
<tr>
<td>BERT +CNN</td>
<td>Lg, cased</td>
<td>Motions</td>
<td>0.27 0.48</td>
<td>0.42 0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manifestos</td>
<td>0.29 0.44</td>
<td>0.54 0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ fine-tune</td>
<td>0.47 0.57</td>
<td>0.61 0.69</td>
</tr>
</tbody>
</table>

6.3.6 Discussion

Through these experiments, I have been able to make a number of observations about the validity and reliability of the annotations produced and the difficulty of the tasks of labelling both debate motions and manifestos.

In labelling the manifestos, the annotators agreed with each other to roughly the same extent that they agree with the gold labels provided by the Manifesto Project’s expert annotators. This level of agreement is also similar to that reported in Mikhaylov et al. (2008), though not as good as that of MARPOR itself (Lacewell and Werner, 2013).

The task does seem to be transferable from manifestos to parliamentary debate motions, with the inter-annotator agreement scores comparable on both.
Although automatic labelling with lexical similarity matching is more successful at the quasi-sentence level than at the motion level, the annotators do not seem to find the coarser grained motion-level task much easier.

Overall, this is a hard task for humans. However, despite the issue of annotation reproducibility, political scientists continue to find these labels useful—as evidenced by Volkens et al. (2015), who found 230 articles that use this data in the eight journals they examined. With comparable reliability (inter-annotator agreement), the labelled motions could prove equally suitable for many automatic analysis applications.

Concerning automation of the labeling process, it is possible to derive three general findings. The first is that a very simple approach—matching debate motions to coded manifestos using cosine similarity measurement—appears to produce potentially useful outputs, particularly at the domain level, with supervised baselines not necessarily offering consistently better results (especially the CNN architectures). The second is that cross-domain applications (from manifestos to motions) seem to necessitate a further fine-tuning step, perhaps due to the very different styles of communication involved. The third is the significant contribution that the use of BERT provides the supervised pipelines, which are able to achieve state-of-the-art classification performance on both the motions and manifesto quasi-sentences.

6.4 Chapter summary

In this chapter, I have explored the task of topic identification in parliamentary debate motions. In Section 6.1.2, I established that topic modelling produces broad, generic topics that are unsuitable for the aims for this project. I therefore took a supervised learning approach, testing hypothesis H4. To this end, I created corpora annotated with labels from two pre-existing sources, and performed classification experiments on these datasets.

In Section 6.2 I evaluated the use of crowdsourced labels from the PublicWhip website. I concluded that, while reasonable performance was observed, this labelling scheme is not scalable to larger and future datasets.

In Section 6.3.4 I evaluated the use of the MARPOR coding scheme and labelled manifesto data for the labelling and automatic identification of the policy preferences expressed in debate motions. I found that classification performance
is also promising using this framework, which offers greater and more systematic coverage of the opinion-topics that appear in motions.

While the use of more complex neural networks alone did not lead to performance gains over SVMs, combining a CNN with BERT contextual embeddings fine trained on data from the target domain did lead to considerable improvements, lending support to hypothesis H4.

The policy preference labels appear to encode both topics and opinions of debate motions in a manner that MPs may respond to in their speeches, with either support or opposition. This chapter has therefore paved the way for analysis of the stance of the individual debate speeches in relation to the policy preferences presented for discussion in the motions. This is the focus of the experiments I present in Chapter 7.
Chapter 7

Topic-centric stance analysis

‘There is nothing new in the compression, manipulation and violation to which political language is subject. The trick is to spot the connections in the network of words.’

Matthew d’Ancona, journalist

In this chapter, I combine the work developed in Chapters 5 and 6 to evaluate systems designed to carry out a form of topic-centric stance analysis. Specifically, they aim to determine the stance of speakers towards the policy preferences proposed in debate motions. For this, I investigate the performances of a range of approaches to debate structure modelling, text representation, and machine learning methods and paradigms. I continue to test hypothesis H4, comparing the performance of contextual word embedding and ANN-based systems against more traditional approaches to text representation and classification.

7.1 Determining support or opposition for policy preferences

As seen in Chapters 3, 5, and 6, previous work in this domain has focused on: (a) sentiment polarity classification (e.g., Bhavan et al. 2019; Burfoot et al. 2011; Thomas et al. 2006); and (b) policy identification (Chapter 6 of this thesis). As far as I am aware, these two tasks have not previously been combined despite the fact that: (1) the information yielded is complementary, and perhaps
even necessary, for practical use (without analysis of debated policies, the target of sentiment in the speeches is unknown); and (2) these two tasks rely on features derived from shared information, which could assist with the learning of parameters for both tasks in a multi-task learning setting. Multi-task learning approaches have been taken to many [NLP] tasks, including part-of-speech tagging, chunking, and named entity recognition (Collobert and Weston [2008]). While such approaches have been successfully applied to sentiment classification of customer reviews (Yu and Jiang [2016]), I am not aware of any previous applications of multi-task learning in the legislative debate domain. In this chapter, I add comparison of single- and multi-task learning paradigms to the text representation and machine learning system components to be evaluated on the task. Figure 7.1 shows frameworks for these two approaches:

Figure 7.1: Single and multi-task learning paradigms for classification of policy preference-focused speaker stance.
Research objectives  The experiments in this section focus on the following objectives:

1. Assessment of the utility of combining the supervised classification of policy preferences and sentiment polarity in order to determine the policy-focused stance of debate speakers.

2. Evaluation of approaches to modelling the policy-focused stance detection task, comparing different learning paradigms, text representations, machine learning methods, and debate models.

7.1.1 Data

In Section 6.3, I described the construction of a manually annotated dataset of policy preferences in debate motions. Unfortunately, that corpus is unsuitable for the task undertaken in this chapter because: (1) it does not include speeches made in response to the motions; and (2) having been collected with manual annotation in mind, the motions in the dataset are all substantive—that is, they ‘express an opinion about something’ (Rogers and Walters 2015), and tend to be of a highly partisan nature, leading to debates in which the stance of MPs can be trivially predicted from their party affiliations. For this study, a mixture of motion types, more representative of the Hansard record as a whole, is required. Moreover, while in Section 6.3, I was able to use textual features derived from the motions themselves, many of the motions in Hansard—and in the corpus I constructed for this study—contain little in the way of informative textual content that could point to the policy preferences under debate. Example 7.1 is typical:

\[ I \text{ beg to move, That the Bill be now read a second time.} \]  

Rather than the content of the motions, for this task, I therefore relied on features derived from the response speeches, which I used as input for the classification of both motions and speeches. For this purpose, I use and adapt the ParlVote corpus developed in Section 5.2. It consists of debates that feature a variety of motion types, and is therefore more representative of the Hansard record as a whole.

1See Section 2.1 for further explanation of debate and motion types.
CHAPTER 7. TOPIC-CENTRIC STANCE ANALYSIS

Annotation

In addition to the vote-derived speech sentiment polarity labels of the original ParlVote corpus, I added policy preference labels for each motion. For this, I annotated each debate motion using annotation guidelines based on those developed in Section 6.3. I revised these to include the new codes used in the updated MARPOR Coding Scheme version 5 (Werner et al., 2015). In this newer version, in addition to the 57 codes used in Section 6.3, twelve of these coding categories have been divided into multiple sub-codes that ‘capture specific aspects of these categories’, making 75 codes in total. I annotated each debate motion following the adapted version of the annotation guidelines, which are available in Appendix E.

As the original ParlVote dataset had not been designed with policy preference analysis in mind, it includes a number of debate types that fall outside the coding scheme. In order to exclude these, I included in the guidelines new instructions to code examples featuring the following types of motions with the label 000: No meaningful category applies:

- **Business of the House** motions, Programme motions, other timetabling and procedural motions, and motions to sit in private. Although it can be possible to interpret such motions as expressing policy preferences, such usage is difficult for an annotator to surmise. Taken at face value, they are concerned simply with the running of Parliament, rather than policy, and were therefore labelled 000.

- Debates with divisions that are not held on the motion in question. In many cases, the division held at the end of the debate concerns some other point that has been brought up during proceedings, such as an amendment. These cases are difficult to exclude automatically, and were therefore coded with 000 during annotation.

- Motions that appear to fit several codes, such as Finance Bills, Local Finance Bills, and Bills concerning the budgets of, for example, police forces. These type of debates tend to focus on multiple and/or local issues, and do not easily fit the MARPOR scheme. Within the area of budgetary Bills is

the exception of motions concerning approval of European Union (EU) Finance Bills, which tend to be either positive or negative towards the EU hence can be labelled 108/110.

- Motions concerning constituency boundary changes, which do not fit the coding scheme.

I excluded all examples given this label from the dataset used for the experiments reported below, as the range of topics covered is likely too broad to constitute a cohesive class. While 56 of the policy preference codes were used as labels by the annotators, I also excluded all examples with policy preference codes that occur fewer than 100 times in the dataset, leaving 34 codes in the data subset used for classification experiments. This left 23,181 example speeches given by 1,321 unique MPs taken from 1,215 different debates. Each example has a manually annotated policy preference label and a vote-derived speech sentiment polarity label. For instance, the example provided in Figure 1.1 is labelled as follows:

Motion: I beg to move, That the Bill be now read a Second time.

Policy preference: 110: European Union: Negative

Speech: I absolutely recognise that people who voted for Brexit did not necessarily vote on economic lines. However, the Government are refusing to publish an impact assessment of this deal. The Prime Minister is expecting MPs to vote for something that we know will damage this country economically, without revealing the impact assessment. What do this Government have to hide?

Sentiment: 0 (negative)

Stance: Opposition to 110

The most common policy preference label in the dataset is 305.1: Political Authority: Party Competence, with 4,926 labelled examples (see Table 7.1)
Table 7.1: Number of examples in the dataset labelled with each MARPOR policy preference code used. Codes used as class labels in the classification experiments described in Section 7.1.2 are highlighted in bold text.

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>n</th>
<th>Code</th>
<th>Name</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>No meaningful category</td>
<td>9524</td>
<td>407</td>
<td>Protectionism: Neg.</td>
<td>43</td>
</tr>
<tr>
<td>101</td>
<td>Foreign Relationships: Pos.</td>
<td>48</td>
<td>411</td>
<td>Technology: Pos.</td>
<td>137</td>
</tr>
<tr>
<td>102</td>
<td>Foreign Relationships: Neg.</td>
<td>12</td>
<td>413</td>
<td>Nationalisation</td>
<td>254</td>
</tr>
<tr>
<td>104</td>
<td>Military: Pos.</td>
<td>398</td>
<td>414</td>
<td>Economic Orthodoxy</td>
<td>54</td>
</tr>
<tr>
<td>105</td>
<td>Military: Neg.</td>
<td>181</td>
<td>416.2</td>
<td>Sustainability: Pos.</td>
<td>13</td>
</tr>
<tr>
<td>106</td>
<td>Peace</td>
<td>155</td>
<td>501</td>
<td>Environ. Protection</td>
<td>631</td>
</tr>
<tr>
<td>107</td>
<td>Internationalism: Positive</td>
<td>67</td>
<td>502</td>
<td>Culture: Positive</td>
<td>14</td>
</tr>
<tr>
<td>108</td>
<td>European Union: Pos.</td>
<td>1601</td>
<td>503</td>
<td>Equality: Positive</td>
<td>1336</td>
</tr>
<tr>
<td>109</td>
<td>Internationalism: Neg.</td>
<td>13</td>
<td>504</td>
<td>Welfare State Expansion</td>
<td>1410</td>
</tr>
<tr>
<td>110</td>
<td>European Union: Neg.</td>
<td>1063</td>
<td>505</td>
<td>Welfare State Limitation</td>
<td>976</td>
</tr>
<tr>
<td>201.2</td>
<td>Human Rights</td>
<td>469</td>
<td>506</td>
<td>Education Expansion</td>
<td>269</td>
</tr>
<tr>
<td>202.2</td>
<td>Democracy—General: Pos.</td>
<td>3</td>
<td>507</td>
<td>Education Limitation</td>
<td>404</td>
</tr>
<tr>
<td>202.3</td>
<td>Repr. Democracy: Pos.</td>
<td>1</td>
<td>601.1</td>
<td>National Way of Life: Pos.</td>
<td>11</td>
</tr>
<tr>
<td>202.4</td>
<td>Direct Democracy: Pos.</td>
<td>166</td>
<td>601.2</td>
<td>Immigration: Neg.</td>
<td>198</td>
</tr>
<tr>
<td>203</td>
<td>Constitutionalism: Pos.</td>
<td>144</td>
<td>602.2</td>
<td>Immigration: Pos.</td>
<td>173</td>
</tr>
<tr>
<td>204</td>
<td>Constitutionalism: Neg.</td>
<td>437</td>
<td>603</td>
<td>Traditional Morality: Pos.</td>
<td>326</td>
</tr>
<tr>
<td>301</td>
<td>Decentralisation: Pos.</td>
<td>570</td>
<td>604</td>
<td>Traditional Morality: Neg.</td>
<td>527</td>
</tr>
<tr>
<td>302</td>
<td>Centralisation: Pos.</td>
<td>398</td>
<td>605.1</td>
<td>Law and Order: Pos.</td>
<td>1399</td>
</tr>
<tr>
<td>303</td>
<td>Govt. and Admin. Efficiency</td>
<td>59</td>
<td>605.2</td>
<td>Law and Order: Neg.</td>
<td>602</td>
</tr>
<tr>
<td>304</td>
<td>Political Corruption</td>
<td>276</td>
<td>606.1</td>
<td>Civic Mindedness: Pos.</td>
<td>11</td>
</tr>
<tr>
<td>305.1</td>
<td>Political Auth.: Party</td>
<td>4926</td>
<td>607.2</td>
<td>Multiculturalism: Pos.</td>
<td>4</td>
</tr>
<tr>
<td>305.2</td>
<td>Political Auth.: Personal</td>
<td>312</td>
<td>608.2</td>
<td>Multiculturalism: Neg.</td>
<td>14</td>
</tr>
<tr>
<td>401</td>
<td>Free Market Economy</td>
<td>1061</td>
<td>701</td>
<td>Labour Groups: Pos.</td>
<td>576</td>
</tr>
<tr>
<td>402</td>
<td>Incentives: Positive</td>
<td>402</td>
<td>702</td>
<td>Labour Groups: Neg.</td>
<td>186</td>
</tr>
<tr>
<td>403</td>
<td>Market Regulation</td>
<td>988</td>
<td>703.1</td>
<td>Agriculture and Farmers: Neg.</td>
<td>25</td>
</tr>
<tr>
<td>405</td>
<td>Corporatism/Mixed Economy</td>
<td>2</td>
<td>705</td>
<td>Middle Class and Prof. Groups</td>
<td>78</td>
</tr>
<tr>
<td>406</td>
<td>Protectionism: Positive</td>
<td>40</td>
<td>706</td>
<td>Underprivileged Min. Groups</td>
<td>230</td>
</tr>
</tbody>
</table>

Inter-annotator agreement

In order to validate the added motion policy preference labels, I recruited a second annotator to label a randomly selected subsection of the corpus. The annotator is an L1 English speaker, with an MA in linguistics and with experience in policy and public affairs. I provide a full data statement, following the guidelines of Bender and Friedman (2018) in Appendix C. After annotation, comparison, and discussion of some initial training examples, she labelled 108 example motions (5.4% of the total).

On this subset, I calculated a Cohen’s kappa agreement score of 0.38, which can be interpreted as representing ‘fair’ agreement (Cohen, 1960). This is comparable to other studies of annotation using the MARPOR codes, including that of Section 6.3, as well as Lacewell and Werner (2013) and Mikhaylov et al. (2008).
This relatively low $\kappa$ value highlights the fact that this is a non-trivial task on which agreement between different human annotators is difficult to achieve. Examination of the confusion matrix of these labels (see Appendix D) reveals that exactly half of all disagreement (31 of 62 instances) concerned label 000, indicating that one of the main challenges of the task may be recognising examples that fall outwith the other codes. Additionally, as many of the labels were used only once, further annotation effort may be required to gain a better understanding of the utility of the labelling scheme.

I make this adapted version of the dataset, ParlVote+, available for the research community.$^3$

### 7.1.2 Method

I investigate approaches to determining, for each example in the dataset, (a) the policy preference expressed in the debate motion (as in Chapter 6), and (b) the sentiment expressed in the speech towards that motion: support (positive) or oppose (negative) (as in Chapter 5). The policy-focused stance of each speaker is then calculated as a combination of these two factors.

For these tasks, I compare the performance of systems comprised of combinations of the following:

- **Learning paradigms** (see Figure 7.1):
  - Single tasks: inputs are processed separately for the two tasks, as in previous work.
  - Multi-task learning: I use a ‘hard parameter sharing’ framework$^{[Ruder(2017)]}$, in which the network shares inputs and parameters in one hidden layer and trains two further task-specific layers separately.

- **Debate models**:
  - Motion-independent: all examples are trained and evaluated together.
  - Motion-dependent: following my findings in Section 5.1, where it appears that that Government-proposed motions tend to be positive and those tabled by opposing parties negative, I trained and evaluated

---

$^3$ParlVote+ corpus available at: [http://dx.doi.org/10.17632/czjfwg9tm](http://dx.doi.org/10.17632/czjfwg9tm), accessed 11 July 2020.
examples from debates initiated by members of the governing and opposition parties separately.

- **Text representations:**
  - BOW: I used tf-idf scores of terms in the dataset to select unigram features (results in Section 5.1 suggest that the addition of higher n-gram features does not improve performance in this domain).
  - Word embeddings: I fine-tuned BERT embeddings on the classification tasks, as in Sections 5.2 and 6.3. As I included uppercase characters in the input, I used the large, cased version of the embeddings.\(^4\)

- **Machine learning classification algorithms.** I used neural networks comprised of two hidden layers. In the multi-task learning setting, with the second of these separated into two task-specific layers (see Figure 7.1). I used a dropout rate of 0.5 for each layer. For binary (speech sentiment) and multiclass (motion policy preference), I used sigmoid and softmax activation layers, respectively. I used early stopping and tested on the model that performed best on the validation set. As in the experiments in Chapters 5 and 6, I compared the following classes of network:
  - MLP: I used networks with hidden layers of 512 nodes.
  - CNN: I used networks with convolutional layers with 512 filters and convolution windows spanning three tokens.

### 7.1.3 Results

I evaluated the systems described above against the majority class for each task. Due to the imbalances in the sets of class labels, I report the macro-weighted F1 score as the evaluation metric.

#### Overall results

Results are presented in Table 7.2. Here, *policy-focused stance* represents the *sentiment polarity* of speakers towards the *policy preference* under debate. I report two measures of this for each system configuration: (1) the mean of the BERT embeddings downloaded from [https://tfhub.dev/google/bert_cased_L-12_H-768_A-12/1](https://tfhub.dev/google/bert_cased_L-12_H-768_A-12/1), accessed 8 July 2020.
F1 scores for policy preference identification and sentiment classification, and (2) the absolute F1 where only examples for which both predicted labels are considered to be correct.

<table>
<thead>
<tr>
<th>Learning paradigm</th>
<th>Text repr.</th>
<th>Machine learning method</th>
<th>Policy pref.</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>Absolute</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ind.</td>
<td>Dep.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ind.</td>
<td>Dep.</td>
</tr>
<tr>
<td>Single-task</td>
<td>— —</td>
<td>Majority class</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>BOW</td>
<td>MLP</td>
<td>0.59</td>
<td><strong>0.65</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.55</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>MLP</td>
<td>0.48</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.38</td>
<td>0.08</td>
</tr>
<tr>
<td>Multi-task</td>
<td>BOW</td>
<td>MLP</td>
<td>0.59</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.59</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>MLP</td>
<td>0.49</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.36</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>— —</td>
<td>Majority class</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>BOW</td>
<td>MLP</td>
<td>0.62</td>
<td><strong>0.67</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>MLP</td>
<td>0.55</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.50</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>BOW</td>
<td>MLP</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.58</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>MLP</td>
<td>0.56</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td>0.48</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Table 7.2: F1 scores for classification of policy preference (multiclass), speech sentiment (binary), and policy-focused stance using motion-independent (Ind.) and motion-dependent (Dep.) debate models. Stance scores are reported as both the mean of the policy preference and sentiment scores and the absolute F1 score. The highest F1 scores for each task are highlighted in bold text.

Most of the tested system configurations comfortably outperformed the naive baselines. In the majority of cases, the motion dependent models performed better than those that did not take into account this aspect of debate structure. In most settings, there was little difference in performance between the two learning paradigms or the two machine learning methods, although the [MLP] performed marginally better than the [CNN] overall (mean F1 scores of 0.53 and 0.51, respectively). [BERT]-based systems tended to perform poorly on policy preference
identification in the motion-dependent setting, perhaps due to the low number of examples per class combined with loss of information due to BERT’s maximum sequence length. The highest overall F1 score for the combined tasks (0.67) was obtained by using single task learning with BOW and MLP in the motion-dependent setting. It is notable that the policy preference detection scores (using BOW) are comparable to those obtained in Section 6.3 despite using completely different input texts, and having no access to the content of the motions themselves.

Results using shorter input speeches

The poorer performance of BERT text representations in all settings is perhaps due to its 512 token sequence input limit. With the mean number of tokens per speech in the ParlVote+ corpus over 700, in many cases, much potentially important information cannot be included when using this framework. Bearing this in mind, in order to test the potential of BERT for this task, I also ran the single task MLP classifier on a subset of the data consisting solely of the 13162 speeches in the dataset that consist of 512 tokens or fewer. Results of these experiments are shown in Table 7.3.

<table>
<thead>
<tr>
<th>Text representation</th>
<th>Policy preference</th>
<th>Speech sentiment</th>
<th>Policy-focused stance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority class</td>
<td>0.01</td>
<td>0.02</td>
<td>0.36</td>
</tr>
<tr>
<td>BOW</td>
<td>0.25</td>
<td>0.40</td>
<td>0.58</td>
</tr>
<tr>
<td>BERT</td>
<td>0.26</td>
<td>0.07</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 7.3: F1 scores for classification of policy preference (multiclass), speech sentiment (binary) and policy stance (mean of these scores) using BOW and BERT-based text representations in the single-task MLP classification setting on speeches of 512 tokens or fewer.

It appears that, for this task, on this corpus, the use of BERT does not lead to the performance gains seen elsewhere. In the motion-independent setting, the two approaches were comparable, while BERT again obtained much lower scores in the motion-dependent setting. Overall, scores were considerably lower than those produced in the main experiments, indicating that, for this task, classification performance benefits from larger quantities of training data, and that this is particularly true here for BERT.
Results by policy preference class

Examining the performance of one of the best performing system configurations—the single-task $\text{BOW}$ $\text{MLP}$ $\text{motion-dependent}$ system—for each (true) policy preference label (Table 7.4), there are a wide variety of scores for each task, with policy preference identification performance ranging from 0.02 to 0.23 and speech sentiment from 0.32 to 1.00.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>0.19</td>
<td>0.44</td>
<td>0.31</td>
<td>411</td>
<td>0.22</td>
<td>1.00</td>
<td>0.61</td>
</tr>
<tr>
<td>105</td>
<td>0.20</td>
<td>0.59</td>
<td>0.39</td>
<td>413</td>
<td>0.12</td>
<td>0.68</td>
<td>0.40</td>
</tr>
<tr>
<td>105</td>
<td>0.23</td>
<td>0.62</td>
<td>0.42</td>
<td>501</td>
<td>0.05</td>
<td>0.62</td>
<td>0.34</td>
</tr>
<tr>
<td>108</td>
<td>0.07</td>
<td>0.63</td>
<td>0.35</td>
<td>503</td>
<td>0.04</td>
<td>0.71</td>
<td>0.38</td>
</tr>
<tr>
<td>110</td>
<td>0.06</td>
<td>0.63</td>
<td>0.35</td>
<td>504</td>
<td>0.09</td>
<td>0.70</td>
<td>0.40</td>
</tr>
<tr>
<td>201</td>
<td>0.06</td>
<td>0.60</td>
<td>0.33</td>
<td>505</td>
<td>0.08</td>
<td>0.66</td>
<td>0.37</td>
</tr>
<tr>
<td>202</td>
<td>0.02</td>
<td>0.46</td>
<td>0.24</td>
<td>506</td>
<td>0.09</td>
<td>0.75</td>
<td>0.42</td>
</tr>
<tr>
<td>203</td>
<td>0.03</td>
<td>0.47</td>
<td>0.25</td>
<td>507</td>
<td>0.06</td>
<td>0.66</td>
<td>0.36</td>
</tr>
<tr>
<td>204</td>
<td>0.11</td>
<td>0.70</td>
<td>0.40</td>
<td>601</td>
<td>0.06</td>
<td>0.49</td>
<td>0.27</td>
</tr>
<tr>
<td>301</td>
<td>0.05</td>
<td>0.63</td>
<td>0.34</td>
<td>602</td>
<td>0.12</td>
<td>0.54</td>
<td>0.33</td>
</tr>
<tr>
<td>302</td>
<td>0.06</td>
<td>0.64</td>
<td>0.35</td>
<td>603</td>
<td>0.08</td>
<td>0.52</td>
<td>0.30</td>
</tr>
<tr>
<td>304</td>
<td>0.10</td>
<td>0.32</td>
<td>0.21</td>
<td>604</td>
<td>0.06</td>
<td>0.63</td>
<td>0.34</td>
</tr>
<tr>
<td>305</td>
<td>0.04</td>
<td>0.76</td>
<td>0.40</td>
<td>605</td>
<td>0.05</td>
<td>0.70</td>
<td>0.38</td>
</tr>
<tr>
<td>305</td>
<td>0.05</td>
<td>0.70</td>
<td>0.38</td>
<td>605</td>
<td>0.06</td>
<td>0.59</td>
<td>0.32</td>
</tr>
<tr>
<td>401</td>
<td>0.04</td>
<td>0.76</td>
<td>0.40</td>
<td>701</td>
<td>0.06</td>
<td>0.76</td>
<td>0.41</td>
</tr>
<tr>
<td>402</td>
<td>0.07</td>
<td>0.60</td>
<td>0.34</td>
<td>702</td>
<td>0.21</td>
<td>0.77</td>
<td>0.49</td>
</tr>
<tr>
<td>403</td>
<td>0.06</td>
<td>0.64</td>
<td>0.35</td>
<td>706</td>
<td>0.12</td>
<td>0.79</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 7.4: F1 scores for policy preference, speech sentiment, and policy-focused stance detection by policy preference code label. The highest score for each task is in bold text. Contrastive pairs of policy preference codes are highlighted in grey boxes.

Each policy preference class received between four and 21 predicted labels in the classifier output ($\mu = 10.4$). Directly contrastive pairs did not necessarily seem to be more difficult to predict than individual class labels, with 104: Military: Positive and 105: Military negative obtaining two of the highest F1 scores. Code 411: Technology and Infrastructure: Positive is in the Economy domain, which contains a number of fairly similar codes. However, this code concerns a well defined topic, and has no directly contrastive partner class, and obtained both perfect sentiment classification and the highest overall F1 score for stance
prediction.

System output analysis

To gain an understanding of the challenges involved in improving classification performance on these tasks, I examined in closer detail the output of the single-task-BOW-MLP-motion-dependent system.

Features of speech polarity  In these experiments, I found that performance was improved by modelling debate structure in the motion-dependent setting. This supports my findings in Section 5.1, in which I observed that the textual features that discriminated between supportive and oppositional speeches were not typically positive or negative when used in the general English language domain.

To investigate how sentiment is manifested in this domain, I first calculated the general-domain sentiment scores of the tokens in each speech example in the test set on a scale of $[-1, 1]$ by looking up the terms in the sentiment lexicon SentiWordNet 3.0 [Baccianella et al., 2010]. These scores are shown in Table 7.5.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>0.44</td>
<td>0.44</td>
<td>0.28</td>
<td>0.25</td>
<td>0.44</td>
<td>0.38</td>
<td>0.44</td>
<td>0.25</td>
<td>0.23</td>
<td>0.44</td>
<td>0.38</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>Min</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.31</td>
<td>-0.38</td>
<td>-0.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Own</th>
<th>Oth.</th>
<th>Own</th>
<th>Oth.</th>
<th>Gov. own</th>
<th>Gov. own</th>
<th>Gov. oth</th>
<th>Gov. oth</th>
<th>Opp. own</th>
<th>Opp. own</th>
<th>Opp. oth</th>
<th>Opp. oth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Own</td>
<td>0.25</td>
<td>0.38</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.21</td>
<td>0.25</td>
<td>0.19</td>
<td>0.25</td>
<td>0.25</td>
<td>0.38</td>
<td>0.14</td>
</tr>
<tr>
<td>Opp.</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.02</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>Other</td>
<td>-0.38</td>
<td>-0.31</td>
<td>-0.38</td>
<td>-0.31</td>
<td>-0.38</td>
<td>-0.31</td>
<td>-0.38</td>
<td>-0.31</td>
<td>-0.38</td>
<td>-0.31</td>
<td>-0.38</td>
<td>-0.31</td>
</tr>
</tbody>
</table>

Table 7.5: Mean sentiment scores for all speeches, supportive (+)/oppositional (-) speeches, replies to Government/opposition party motions, responses to own/other party motions, and all combinations of these three factors.

The mean sentiment of speeches overall is very slightly negative (-0.01), according to the lexicon. Overall however, there is little difference between supportive and oppositional speeches in the polarity of language used. This is also the case for speeches given in different scenarios, such as in response to Government/opposition motions, by speakers addressing motions proposed by members with their own or with different party affiliations, or any combinations of these factors. This demonstrates once again that terms used in parliamentary debate
speeches do not usually express the same sentiments that they may be expected to do in general usage.

<table>
<thead>
<tr>
<th>Motion-independent</th>
<th>Motion-dependent</th>
<th>Opposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Government</td>
<td>Opposition</td>
</tr>
<tr>
<td>Labour</td>
<td>+0.13</td>
<td>+0.17</td>
</tr>
<tr>
<td>Gentleman</td>
<td>+0.13</td>
<td>0.00</td>
</tr>
<tr>
<td>shadow</td>
<td>−0.09</td>
<td>−0.03</td>
</tr>
<tr>
<td>Prime</td>
<td>+0.09</td>
<td>+0.13</td>
</tr>
<tr>
<td>party</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>cuts</td>
<td>+0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Lady</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>situation</td>
<td>−0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>threat</td>
<td>−0.28</td>
<td>−0.03</td>
</tr>
<tr>
<td>outside</td>
<td>0.00</td>
<td>+0.06</td>
</tr>
<tr>
<td>pay</td>
<td>+0.06</td>
<td>+0.19</td>
</tr>
<tr>
<td>Lords</td>
<td>0.00</td>
<td>+0.08</td>
</tr>
<tr>
<td>crisis</td>
<td>−0.06</td>
<td>+0.13</td>
</tr>
<tr>
<td>Government</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>constituents</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>wants</td>
<td>−0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>important</td>
<td>+0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>careful</td>
<td>+0.19</td>
<td>+0.13</td>
</tr>
<tr>
<td>week</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>stop</td>
<td>−0.02</td>
<td>+0.06</td>
</tr>
</tbody>
</table>

Table 7.6: Top 20 discriminating features for the motion-independent setting (all speeches), and, in the motion-dependent setting, responses to Government- and opposition-proposed motions, together with their mean SentiWordNet scores.

To examine which terms in the speeches do indicate sentiment, I obtained the permutation importance scores of each unigram in the input vocabulary. That is, for feature \( j \) in the feature set \( N \), I calculated the permutation feature importance as the difference between performance (in this case, the F1 score) using the original dataset \( D \) and a corrupted version \( \tilde{D} \), in which \( j \) has been randomly shuffled (Breiman, 2001). I consider features with higher scores to be more important to the model. A sample of the most important features (the top 20) in each setting according to this metric is shown in Figure 7.6. Comparing (the lemmas of) these terms with their SentiWordNet scores (means over all word senses), it seems that the features that are indicative of support or opposition are not those that would typically be used for subjective expression in
general English usage. Rather, many are parliamentary terms, such as forms of address, and other proper nouns. This is particularly true for speeches addressing opposition-proposed motions.

**Party affiliations** As MPs usually vote along party lines (see Subsection 5.1.1), it would be possible to achieve good sentiment classification results by setting a classifier to make predictions on that simple basis. On the other hand, we also know that MPs are more free to ‘rebel’ against their parties in their speeches than in their voting behaviour (see Section 5.1.1). To investigate how this effects sentiment polarity classification, I compared the performance of rebel MPs—those voting against a motion proposed by their own party or in support of one proposed by another party—and loyal MPs. This produced F1 scores of 0.77 and 0.66 respectively. The lower performance on loyal voters may suggest that, on occasion, speakers may use language that goes some way towards supporting the position of their opponents, while ultimately voting with their parties, and that these cases may be harder to detect than outright rebellions.

The frequency distribution plots in Figure 7.2 present a closer look at this. They show the predicted probabilities of examples being assigned to the positive class. I compare the probability distributions for correctly and incorrectly predicted testset examples. These densities are shown in three settings: all predictions, intra-party speeches (made in response to motions proposed by an MP with the same party affiliation), and inter-party responses (replies to a member of another party).
There are a number of clear patterns in the distributions. Overall, the system tends to make more confident predictions for examples that it predicts correctly (that is, it outputs probabilities towards 0.0 for negative and 1.0 for positive examples), and is less confident about examples that it predicts incorrectly (closer to 0.5), as might be expected. In the intra-party setting, the model outputs high probabilities that it assigns to the positive class (correctly, more often than not). Meanwhile, negative predictions (usually incorrect) are made with probabilities that tend towards 0.5 (that is, with low certainty). For inter-party response speeches, this pattern is reversed, albeit not to as dramatic an extent. This may be due to situations in which, for example, multiple opposition parties collaborate against the Government, which introduce some noise into this analysis. Ultimately, the patterns seen here suggest that the language used in the speeches may often say more about the speakers’ party affiliations than it does about the nuances of individual speaker stance.

**Input speech length** The length of speeches does not seem to greatly affect classification, with examples that are classified correctly, partially correctly, and completely incorrectly having similar distributions of token numbers (see Table 7.7).

<table>
<thead>
<tr>
<th></th>
<th>Stance ✓</th>
<th>PP ✓, sent. ✓</th>
<th>Sent. ✓, PP ✓</th>
<th>Stance ✓</th>
</tr>
</thead>
<tbody>
<tr>
<td>n examples</td>
<td>1120</td>
<td>404</td>
<td>492</td>
<td>303</td>
</tr>
<tr>
<td>Max. tokens</td>
<td>20730</td>
<td>6505</td>
<td>6484</td>
<td>4742</td>
</tr>
<tr>
<td>Mean tokens</td>
<td>876.9</td>
<td>916.5</td>
<td>761.4</td>
<td>867.6</td>
</tr>
<tr>
<td>Min. tokens</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Std. deviation</td>
<td>1213.9</td>
<td>953.2</td>
<td>1115.6</td>
<td>4742</td>
</tr>
<tr>
<td>&lt; 50 tokens</td>
<td>103</td>
<td>43</td>
<td>61</td>
<td>35</td>
</tr>
<tr>
<td>&gt;= 50 tokens</td>
<td>1017</td>
<td>361</td>
<td>431</td>
<td>268</td>
</tr>
</tbody>
</table>

Table 7.7: Number of speeches by token counts and prediction outcome (√ = correct and X = incorrect).

While some previous work has excluded speeches of fewer than 50 tokens under the assumption that they are unlikely to contain enough information to express sentiment (Salah (2014), Section 5.1), it appears that this may be unnecessary. In the experiments, 67.8% of these shorter examples were classified correctly for speech sentiment (compared with 69.5% of examples of any length), and 42.6% of examples < 50 classified correctly on both tasks (48.1% for the whole dataset). With examples of both very short speeches (such as two-word speeches like ‘Hear
hearth', ‘Under Labour’) and the longest speech examples classified correctly, it seems that speech length is not an important factor in performance for the BOW-based systems.

7.1.4 Discussion

Policy-focused stance detection of parliamentary speeches is a challenging task, which I have framed as combined multiclass and binary classification of policy preferences and sentiment, respectively. However, despite the large number of classes in the labelling scheme for the policy prediction task, and the fact that the input features used were based only on the content of speeches (not the motions or titles, as in previous work), the tested systems were able to obtain potentially useful results, beating naive baselines.

Modeling of the structure of parliamentary debates in the form of motion-dependent classification had been seen to improve performance on speech sentiment classification in Chapter 5. In the experiments described in this chapter, I found that this model is not only consistently superior for speech sentiment classification, but also improves the identification of policy preferences—the topics under discussion. I have shown that the differences between supportive and opposing speeches do not tend to derive from terms that typically express sentiment in general English language domains, but from the relationships between the speaker, the MP who proposes the motion in question, and the party affiliations of both actors.

The application of multi-task learning did not, in most configurations, improve system performances. However, I used a fairly simple framework in which just one of the network’s hidden layers was shared with one further hidden layer per classification task. There is therefore scope for further experimentation with this approach.

In these experiments, fine-tuning on BERT embeddings led to considerably worse performances in policy preference identification. Considering the successes of this approach on other tasks and domains, this also warrants further investigation. It may be that, rather than fine-tuning on domain embeddings, a more domain-specific approach is desirable in this context.
7.2 Chapter summary

In this chapter, I have combined the tasks explored previously in this thesis to formulate the task of policy preference-focused stance detection. To this end, I adapted the corpus constructed in Section 5.2.1 to include manually annotated policy preference labels, and evaluated the performances of combinations of a range of approaches to text representation, debate structure modelling, and machine learning paradigms and methods. Analysis of the output of these systems has provided insights into the nature of the parliamentary debate domain and its challenges.

The results presented here reinforce support for hypothesis H2, showing that in most settings, stance detection systems benefit from modelling the polarity shifting effects of debate motions. However, hypothesis H4 is not supported by the results, suggesting that state-of-the-art approaches such as BERT may require further domain-specific adaptation. Overall, I have shown that the task of policy-focused stance detection in debate speeches can be feasibly automated—even with simple features and neural architectures—in both single and multi-task settings.
Part III

Conclusions
Chapter 8

Conclusion

‘If politics is, on occasions, the place of low skulduggery, it is more often the place for the pursuit of noble causes. I wish everyone, friend or foe, well. That is that. The end.

Tony Blair, Prime Minister 1997-2007

In this concluding chapter, I revisit the research objectives and questions that I established in Section 1.3, and assess the extent to which I have fulfilled and answered them. I also propose ideas and recommendations for extending this work in the future.

Evaluation of research objectives

Initially, in order to understand the current state of sentiment analysis in the domain of parliamentary and legislative debates, I had posed research question RQ1 (What approaches have been taken to the automatic analysis of speaker’s sentiment and position-taking in parliamentary and legislative debates? What challenges does this domain pose for the application of sentiment analysis?). To answer this, I set objective O1:

To establish the current state of approaches to sentiment analysis of parliamentary and legislative debates by conducting a systematic review.

In Chapter 3 I fulfilled this by completing an extensive survey of sentiment analysis in this domain by researchers from diverse fields. This constitutes my
research contribution C2: A systematic review of literature concerning sentiment and position taking analysis of parliamentary and legislative debates. From the findings of this review, I determined that: (1) generally held assumptions regarding the validity of general-domain approaches to sentiment analysis had not been challenged in prior work; (2) for sentiment classification in this domain, no attempts had previously been made to determine the nature of the targets of sentiment, the opinion-topics or policies under discussion; (3) existing approaches tended not to combine potentially beneficial knowledge and theory from different research fields such as political science and computer science; and (4) recent developments in NLP that have achieved state-of-the-art results in other settings (such as transformer-based contextual word embeddings) had not been tested in this domain. These findings informed my approach to the empirical evaluations conducted in Part II.

To gain an understanding of the nature of the task at hand, I established objective O2:

To assess the available corpora and data, and identify the particular characteristics of UK parliamentary debates.

To meet this objective, in Chapter 2 I described the existing transcript data, as well as the format and structure of parliamentary debates. This information supported me in fulfilling the next objective O3:

To construct annotated corpora and datasets for use in the evaluation of systems designed for sentiment polarity classification, multi-class topic classification, and topic-centric sentiment analysis of UK parliamentary debates.

For the experiments conducted in Part II of the thesis, I constructed annotated corpora designed for the evaluation of each task. These datasets constitute my research contribution C1 (new annotated corpora for the evaluation of systems designed for sentiment polarity classification, opinion-topic identification, and topic-centric sentiment analysis of UK parliamentary debates.) I have made these corpora publicly available, which has resulted in further investigation of the problem by other researchers (Bhavan et al., 2019; Sawhney et al., 2020).
These corpora then constitute the input data for the empirical analyses that I conducted in Part I in order to fulfill objective O4:

To design and evaluate approaches and methods for (1) sentiment polarity classification, (2) topic identification, and (3) topic-centric sentiment analysis of UK parliamentary debate speeches that take account of and exploit the structure and characteristics of the documents in this domain.

This objective aimed to answer the following three research questions:

**RQ2: Sentiment polarity classification** How effective are general approaches to sentiment polarity classification when applied to the domain of UK parliamentary debates? What characteristics of the debates effect the performance of such systems? How can sentiment classification be optimized for this domain?

In addressing this research question, I made contribution C3: Evaluation of approaches to sentiment polarity classification optimized for the domain of UK parliamentary debate speeches.

In Chapter 5 I examined hypothesis H1: MPs’ votes are unreliable sentiment/stance class labels, as they do not always reflect the opinions and positions expressed in their speeches. I found that division vote labels do not always represent sentiment polarity as recognised by human readers. While the costs associated with manual annotation have prohibited extended investigation, the relationship between ground-truth sentiment and its representation in automatic systems is certainly worthy of further study. This is particularly the case for examples of ‘rebel’ speech, which are not necessarily well represented by the voting record, but are likely to be among the most interesting examples for creators and users of automated analysis systems.

In Chapters 5 and 7 I found that language appears to be used differently in Parliament than in the general domain, with different types of words indicative of positive and negative sentiment. Moreover, I found that the structure of the debates has significant effects on the polarity of lexical terms, which can have serious consequences for the effectiveness of automatic sentiment classifiers. I tested research hypothesis H2: The polarity shifts caused by the discourse structure of parliamentary debates can be mitigated by applying a two-step, debate motion-dependent classification model. I found that both the two-step classification and
Government/opposition labelling models led to substantial performance gains. These results suggest that the classification methods that are typically applied in other domains can be enhanced by taking this aspect of debate structure into account.

However, I also found that the application of BERT contextual embeddings does not seem to lead to improvements in system performance on this task, perhaps due to the large length of most of the input speeches. Considering the successes achieved with BERT and BERT-like approaches elsewhere in NLP adapting them to take into account the nature of this task and domain is an avenue for exploration in future research.

**RQ3: Opinion-topic identification** What are the characteristics of the topics that are the targets of sentiment expressed in UK parliamentary debate speeches. How can these be labelled and automatically identified?

Finding that unsupervised topic-modelling approaches are unlikely to produce the type of opinion-topics that are targets of subjective debate speech (Section 6.1.2), in Chapters 6 and 7, I examined hypothesis H3: Supervised machine learning classifiers can predict the manually applied opinion-topic labels of debates from textual (and metadata) features of the debate motions and speeches in a multiclass setting.

I reframed the topics as policies or policy preferences, and tested the utility of two pre-existing labelling schemes, one crowd-sourced, and the other developed by experts in political science. I found that, while it is possible to achieve reasonable document classification results, there are limitations of scale and annotation cost associated with the two labelling schemes, respectively. To mitigate the latter, future work could investigate ways of improving performance of the unsupervised similarity-matching approach applied in Section 6.3.4.

Unlike the sentiment classification experiments and the policy preference identification element of Chapter 7, here use of BERT-based text representation did boost classification performance with the shorter-form input of debate motion texts. There is however, further scope for experimentation with the design of more complex neural architectures for this task, which I leave for future work.

These results represent contribution C4: Evaluation of labelling schema and approaches to the identification of the policy preferences (opinion-topics) discussed in UK parliamentary debate motions (proposals).
RQ4: Topic-centric sentiment analysis What approaches and methods to topic-centric sentiment analysis are most effective for detecting the policy preferences of speakers in UK parliamentary debates? What are the remaining challenges associated with this task?

Here, I assessed the effectiveness of combining the approaches taken to sentiment and topic classification outlined above, and made research contribution C5: Evaluation of approaches to the novel task of policy preference-focused stance detection—a form of topic-centric sentiment analysis—in the domain of UK parliamentary debates.

These experiments demonstrated that the task of policy-focused stance detection in debate speeches can be automated, even with fairly simple features and neural architectures. However, as the proposed approaches that exploit BERT embeddings and a multi-task learning paradigm did not lead to performance gains, optimizing these methods for this task and domain could be a direction for future work.

Analysis of the output of the classification system revealed a number of insights regarding the nature of the task. It is possible to identify the policy preference label of debate motions without providing the system with access to the content of the motion itself. While long speeches may not be suitable for BERT-based text representation, the classifiers seem to be able to handle very short speeches, which had been assumed not to be the case in prior work. Sentiment polarity appears to be easier to predict for rebel MPs than loyal ones, but ultimately, the classifier may recognise party affiliations more easily than the nuances of policy positions. A possible avenue for exploration in future work is to take a closer look at the relationship between MPs, their voting behaviour, the language of their speeches, and the ground-truth of speaker stance.

Overall, the experimental results presented in Part II are mixed with regards to hypothesis H4: Classification performance on sentiment and stance detection of parliamentary debate speeches benefits from approaches to text representation and machine learning paradigm that have achieved state-of-the-art results in other domains, specifically contextual embeddings and artificial ANN machine learning methods. In most settings, the use of neural networks and more complex text representations does not appear to produce substantial benefits for classification in this domain over more traditional methods. A potential direction for further
development of this work is to adapt these methods for the parliamentary debates domain by, for example, pre-training language models on in-domain data.

In this thesis, I have framed the task of determining the positions taken in parliamentary speeches as one that necessarily makes use of, and is reliant on, information relating to the topics discussed in debates. For this, I have combined sentiment classification with an approach to topic labelling informed by political science. While the focus of this thesis has been debates held in the UK Parliament, the approaches taken here could be extended to other legislatures, particularly those based on the Westminster system, such as the Australian, Canadian, and New Zealand parliaments. The work presented here may represent a step in the development of NLP systems for democracy-focused civic technology that would allow people to more-easily monitor the activities of their elected representatives.
Bibliography


A. Balahur, Z. Kozareva, and A. Montoyo. Determining the polarity and source of opinions expressed in political debates. In G. A., editor, Computational


BIBLIOGRAPHY


BIBLIOGRAPHY


R. Sawhney, A. Wadhwa, S. Agarwal, and R. R. Shah. GPolS: A contextual graph-based language model for analyzing parliamentary debates and


Appendix A

HanDeSeT annotation guidelines

A.1 Introduction

The purpose of this annotation task is to identify the sentiment polarity of motions (proposals put to the assembly) towards the subjects discussed in Parliamentary debates, and to identify the sentiment polarity expressed by speakers in those debates towards these motions.

This task uses transcripts of debates from the House of Commons of the UK Parliament. For each debate, a motion is proposed, and selected members of the assembly respond one or more times to the motion and to the utterances (speech segments) of their fellow speakers. For this task, annotators are presented with both the motion and all the utterances produced by individual speakers in a given debate. The combined utterances of a single speaker (between one and five utterances) are considered to form a speech.

The task of the annotator is twofold:

1. To categorize the motion of each debate as being either broadly positive or broadly negative towards the subject of debate.

2. To determine whether, overall, the stance towards the motion under debate of each speaker’s speech is broadly positive (in support of the motion) or broadly negative (against the motion).
A.2 Annotation procedure

Each unit in this task consists of a debate motion and between one and five utterances that comprise a speaker’s speech. Motions typically address subjects of parliamentary interest (e.g., Government policy in areas such as health or education), Papers, Bills or Acts, aspects of Parliamentary procedure (e.g., the amount of time to be dedicated to a particular debate), or amendments to legislation brought by the Government, an opposition party, or an individual Member.

1. The annotator should read the motion carefully, decide whether it is broadly positive or negative towards the subject of the debate, and assign it the corresponding label: ‘1’ for positive, ‘0’ for negative.

2. In many cases multiple units are extracted from the same debate and therefore have the same motion. The annotator should ensure that (s)he assigns all identical motions the same label.

3. For each unit, the annotator then reads the speaker’s speech comprised of one to five utterances. (S)he considers the overall sentiment polarity of the entire speech, and then assigns a sentiment polarity label to the speech in question. Again, the label ‘1’ signifies positive sentiment in support of the motion, while the label ‘0’ signifies negative sentiment and is against the motion.

A.3 Annotation guidelines for motion and speech sentiment polarity

Based on the sentiment expressed towards the motion in question, each speech is assigned either a positive or negative polarity label. For example:

A.3.1 Annotation examples

i. Motion: That this House approves the draft Agreement (Cm 9332), between the Secretary of State for Culture, Media and Sport and the British Broadcasting Corporation, which was laid before this House on 15 September 2016.
(a) **Speech:**

Utterance 1: *Many Members have expressed the view that the BBC is indeed one of our most beloved cultural institutions. Each of us will have fond memories of the TV shows that made us laugh and cry, and those that educated and inspired us. To this very day, some of the world’s most famous TV programmes call the BBC their home, or can at least trace their roots back to it. The BBC also has a proud record of supporting and cultivating some of Britain’s most treasured personalities and actors. With the BBC’s global reach, all this goes a significant way towards promoting our place in the world. It is perhaps the largest exporter of our cultural values, and it is viewed by hundreds of millions of people. Some might even say it is our best soft power asset. However, domestic and global habits continue to change, and for the BBC’s importance to be maintained, it needs to change with them.*

Utterance 2: *Our BBC is not perfect, and it has long needed action to address governance issues and changing viewing habits. I was pleased that those issues were highlighted by all parties at the start of the 2015 negotiations. Now is the time to see them addressed and for solutions to be approved. Like many hon. Members, I have received a tremendous number of representations from constituents who are concerned about the BBC’s future. Given that our constituents pay a licence fee, our communities have a rightful stake in this institution. I am pleased that the new royal charter has been taken seriously and dealt with positively by the Government.*

Utterance 3: *Under the draft agreement, I see a BBC that suits the modern broadcasting and digital environment that we know today. Much has been said about the new governance structure for the BBC. Since the publication of the White Paper, real progress has been made on the subject of appointments to the BBC board through discussion and consultation with the BBC. The fact that the BBC will appoint a large majority of its board members for the first time is indeed a positive measure that clearly maintains its independence.*

Utterance 4: *It is right that all the nations that make up the United Kingdom are represented on the BBC board and that these individuals are...*
subject to the public appointments process. It is also right that those appointments should not be subject to undue political influence. However, it is right, too, that the Government retain a role in appointing non-executive directors to the board of a body that spends £3.7 billion of public money each year. We are talking about huge sums that have to be justified. We cannot allow waste or a lack of openness when it is the public who have such a sizeable stake, yet with the expanded role of the National Audit Office and Ofcom as overseers of the BBC’s financial and content scrutiny arrangements, I am certain that we will maintain the credibility expected of our public service broadcaster.

Utt. 5: The BBC is a huge part of our past, our present and our future. The new charter and agreement will enable improvements that will ultimately address the important issues of governance and modernisation while ensuring the BBC’s independence and enhancing the distinctiveness of its content. I am therefore pleased to support the motion and agreement, which will guarantee the BBC’s important place in our society for many years to come.

Assigned motion sentiment polarity label = 1 (positive)
Assigned speech sentiment polarity label = 1 (positive)

ii. Motion: That this House regrets the impact of school funding cuts on the ability of children to reach their full potential; and calls on the Government to ensure that all schools have the funding that they need to provide an excellent education for every child.

(a) Speech:

Utt. 1: I thank my hon. Friend for letting me get in at last. Does my hon. Friend agree that it is grossly unfair that the pupils of Somerset have had, on average, £2,000 per pupil less than the national average? We are very grateful to the Government for increasing funding to Taunton Deane by 4.5%. This will make it fair, when historically things have been grossly unfair.

Utt. 2: My hon. Friend is making a very valid point about early-years. Does she agree that this is not just about a new fairer funding formula? This Government are putting much money into education, particularly
for the new 30 hours of free childcare. Neroche pre-school in my con-
stituency is having a brand-new building built on the back of that money
and it is only too grateful to the Government.

Utt. 3: Does my hon. Friend agree that many of our rural schools in
Somerset and Dorset have been doing so well with the funding they
have had? This extra funding might enable them to put in place some
of the things that they have not been able to have because there simply
has not been enough money to go around.

Assigned motion sentiment polarity label = 0 (negative)
Assigned speech sentiment polarity label = 0 (negative)

iii. Motion: Proceedings on consideration shall be taken in the order shown in
the first column of the following Table and shall be brought to a conclusion
(so far as not previously concluded) at the times specified in the second
column.

<table>
<thead>
<tr>
<th>Proceedings.</th>
<th>Time for conclusion of proceedings</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Clauses relating to the purposes and scope of registration and identity cards.</td>
<td>2.45 p.m. on the day on which proceedings on consideration are commenced.</td>
</tr>
<tr>
<td>New Clauses relating to the National Identity Scheme Commissioner, Amendments relating to Clauses 24 and 25.</td>
<td>3.45 p.m. on that day.</td>
</tr>
<tr>
<td>Remaining new Clauses, Amendments relating to Clauses 1 to 23, Amendments relating to Clause 26, Amendments relating to Clauses 33 to 45, new Schedules, Amendments relating to Schedules 1 and 2.</td>
<td>4.30 p.m. on that day.</td>
</tr>
<tr>
<td>Amendments relating to Clauses 27 to 32 and any remaining proceedings on consideration.</td>
<td>5.00 p.m. on that day.</td>
</tr>
</tbody>
</table>
(a) Speech:

Utterance 1: As someone who has made it clear that I am not in favour of the Government’s proposals, having written a minority report for the Home Affairs Committee and voted against the Bill on Second Reading, may I put it to the hon. Gentleman that, whatever the Government’s failure to provide time, we are where we are? If the purpose of the main Opposition party is to utilise all the time allowed to debate the programme motion, that would effectively result in less time to debate the matters of substance before us. It would be most unfortunate if that were its tactic, because it would aggravate a situation in which the Government have not provided the time that we would like.

Assigned motion stance polarity label = 1 (positive)
Assigned stance polarity label = 1 (positive)

(b) Speech:

Utterance 1: I very much agree with my hon. Friend, but I hope that he accepts that because many Conservative Members take such comments seriously, we need a constant reminder that when we are returned to Government automatic guillotining will stop. That has not been stated with sufficient strength, and until my hon. Friend does so his comments lack moral authority.

Utterance 2: Does the hon. Gentleman agree that the other problem is that there are disagreements within parties about such Bills? I happen to be very much in favour of the principle of the Bill, while others in my party take a different view. This is the only opportunity that we all have to show what parts we agree with and where we might want alteration, so that there can be greater consensus. With a Bill of this sort and of this importance, that consensus becomes very important.

Assigned motion sentiment polarity label = 1 (positive)
Assigned sentiment polarity label = 0 (negative)

A.3.2 Explanation of label assignments

i. Motion $i$ is assigned the positive label, 1, as it asks the House to ‘approve’ a piece of legislation.
The speech in example \( i(a) \) is also judged to be positive and assigned the label 1. Extracts that display a positive stance towards the motion include:

- *I am pleased that the new royal charter has been taken seriously and dealt with positively by the Government.*
- *Under the draft agreement, I see a BBC that suits the modern broadcasting and digital environment that we know today.*
- *I am therefore pleased to support the motion and agreement, which will guarantee the BBC’s important place in our society for many years to come.*

ii. Motion \( ii \) is assigned the negative label, 0, as use of the word ‘regret’ expresses dissatisfaction with Government education policy.

Speech \( ii(a) \) is judged to be negative and assigned the label 0. As the motion itself is negative towards actions of the Government (school funding cuts), speeches with a positive stance towards the motion take the opposite position to that of the Government. This speech is supportive of the Government, and therefore exhibits negative sentiment towards the motion.

Extracts that display negative sentiment towards the motion include:

- *We are very grateful to the Government for increasing funding to Taunton Deane by 4.5%.*
- *Neroche pre-school in my constituency is having a brand-new building built on the back of that money and it is only too grateful to the Government.*

iii. Motion \( iii \) is assigned the label 1, as it proposes that an activity take place (the timetable of a debate).

Many motions, such as this one, concern procedural matters related to the discussion of a piece of legislation. In such cases the sentiment of a speech towards the motion may be in contradiction to the sentiment it displays towards the legislation or wider topic in question. In speech \( iii(a) \), the speaker is opposed to the amendments that are to be discussed:

- *I am not in favour of the Government’s proposals.*
- *I voted against the Bill.*
However, the speaker accepts the timetable for the debate proposed in the motion:

- *whatever the Government’s failure to provide time, we are where we are.*
- *to utilise all the time allowed to debate the programme motion ... would aggravate a situation in which the Government have not provided the time that we would like.*

This speech is therefore assigned a positive label (1).

The sentiment of speech *iii(a)*, however, is in favour of the legislation in question:

- *I happen to be very much in favour of the principle of the Bill.*

But it exhibits negative sentiment towards the motion (which the speaker considers to be ‘guillotining’ of the debate):

- *when we are returned to Government automatic guillotining will stop.*
- *This is the only opportunity that we all have to show what parts we agree with and where we might want alteration.*

It is therefore assigned a negative label (0).

### A.3.3 Dealing with ambiguity in motion stance

Motions may contain both positive and negative language. It is often the case that a motion expresses disapproval of a situation or policy (negative), yet calls for some form of action (positive). Additionally, motions may include praise and/or condemnation of various people, groups, organisations etc. further complicating the task.

The following examples illustrate correct sentiment polarity labelling for such motions:

---

1Guillotine: ‘a procedure used to prevent delay in the discussion of a legislative bill by fixing times at which various parts of it must be voted on.’ OED
i. Motion: That this House regrets the impact of school funding cuts on the ability of children to reach their full potential; and calls on the Government to ensure that all schools have the funding that they need to provide an excellent education for every child.

Assigned motion sentiment polarity label = 0 (negative)

ii. Motion: That this House recognises the contribution of student nurses, midwives, allied health professionals and other healthcare staff; has serious concerns about the potential impact of removing NHS bursaries on the recruitment and retention of staff; and calls on the Government to drop their plans to remove NHS bursaries and instead to consult on how they can best fund and support the future healthcare workforce.

Assigned motion sentiment polarity label = 0 (negative)

Explanation: In both these cases, the motion requests action from the Government (positive action in i, negative in ii). However, the overall tone of both motions is negative towards the Government’s policies, and they should therefore be labelled as expressing negative sentiment.
Appendix B

Debate motion annotation guidelines

B.1 Introduction

This project aims to use codes developed by the Comparative Manifesto Project (CMP) for the annotation of political parties’ manifestos (election programmes) to label the policy positions expressed by Members of the United Kingdom Parliament (MPs) in debate motions.

In the House of Commons (the lower chamber of Parliament), each debate normally begins with a motion, a proposal tabled by an MP. These motions express the MPs’ (and usually their parties’) policy positions on one or more topics.

The unit of analysis is the quasi-sentence, a sentence or sentence clause, that should contain exactly one statement or “message”. For each quasi-sentence, the aim is to find the most suitable code.

B.2 Policies and coding

The coding scheme is the Manifesto Project version four (mp v4), and is available at https://manifestoproject.wzb.eu/coding_schemes/mp_v4.

There are 57 codes in the scheme, each of which consists of three digits. 56 of these categories are grouped into seven major policy domains. The ‘hundreds’ position in each signifies this category, while the ‘tens’ and ‘units’ positions indicate the specific policy preference.
• Domain 1: External Relations
• Domain 2: Freedom and Democracy
• Domain 3: Political System
• Domain 4: Economy
• Domain 5: Welfare and Quality of Life
• Domain 6: Fabric of Society
• Domain 7: Social Groups

The 57th code is ‘000 No meaningful category applies’, and is used for ‘statements not covered by other categories; sentences devoid of any meaning.’

Further information and examples of coded manifesto excerpts are in the CMP coding Handbook, available at: https://manifestoproject.wzb.eu/down/papers/handbook_2011_version_4.pdf The following notes are to be used in conjunction with the coding scheme:

<table>
<thead>
<tr>
<th>Code</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Foreign Special Relationships: Positive</td>
</tr>
<tr>
<td>102</td>
<td>Foreign Special Relationships: Negative</td>
</tr>
<tr>
<td>103</td>
<td>Anti-Imperialism</td>
</tr>
<tr>
<td>104</td>
<td>Military: Positive</td>
</tr>
<tr>
<td>106</td>
<td>Peace</td>
</tr>
<tr>
<td>Code</td>
<td>Freedom and Human Rights</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>202</td>
<td>Democracy</td>
</tr>
<tr>
<td>302</td>
<td>Centralisation</td>
</tr>
<tr>
<td>305</td>
<td>Political Authority</td>
</tr>
<tr>
<td>402</td>
<td>Incentives</td>
</tr>
<tr>
<td>404</td>
<td>Economic Planning</td>
</tr>
<tr>
<td>405</td>
<td>Corporatism/Mixed Economy Positive</td>
</tr>
<tr>
<td>406</td>
<td>Protectionism: Positive</td>
</tr>
<tr>
<td>408</td>
<td>Economic Goals</td>
</tr>
<tr>
<td>409</td>
<td>Keynesian Demand Manage ment</td>
</tr>
<tr>
<td>415</td>
<td>Marxist Analysis: Positive</td>
</tr>
<tr>
<td>503</td>
<td>Equality: Positive</td>
</tr>
<tr>
<td>504</td>
<td>Welfare State Expansion</td>
</tr>
<tr>
<td>Code</td>
<td>Category</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>606</td>
<td>Civic Mindedness: Positive</td>
</tr>
<tr>
<td>607</td>
<td>Multiculturalism: Positive</td>
</tr>
<tr>
<td>608</td>
<td>Multiculturalism: Negative</td>
</tr>
<tr>
<td>704</td>
<td>Middle Class and Professional Groups</td>
</tr>
<tr>
<td>705</td>
<td>Underprivileged Minority Groups</td>
</tr>
<tr>
<td>706</td>
<td>on-economic Demographic Groups</td>
</tr>
<tr>
<td>000</td>
<td>No meaningful category applies</td>
</tr>
</tbody>
</table>
Appendix C

Data statement for ParlVote+

Data set name: ParlVote+

Data set developers: Gavin Abercrombie

Data statement author: Gavin Abercrombie

Others who contributed to this document: Nancy Greig

A. Curation rationale

A dataset for policy-focused stance detection in United Kingdom (UK) parliamentary debates. This dataset is an adaptation of ParlVote: A Corpus for Sentiment Analysis of Political Debates with policy preference labels added to each example.

B. Language variety/varieties

BCP-47 language tag: en-GB

Language variety description: Near-verbatim transcribed standard British English as spoken in the House of Commons of the UK Parliament.

C. Speaker demographic

Description: Members of the United Kingdom Parliament (MPs)

Number of different speakers represented: 1,321

D. Annotator demographic

Description: One primary and one secondary annotator

Age: 43, 42

Gender: M, F

Race/ethnicity: white British/Scottish

First language(s): English

Training in linguistics/other relevant discipline: MSc Cognition, MA Linguistics

E. Speech situation

Description: parliamentary debate

Time: 1997-2019

Place: House of Commons, UK Parliament

Modality (spoken/signed, written): transcribed

Scripted/edited vs. spontaneous: mixed

Synchronous vs. asynchronous interaction: asynchronous

Intended audience: other MPs, the media, the general public
F. Text characteristics

Debates are moderated by the Speaker (chief presiding officer of the House) following parliamentary rules of behaviour\(^2\)

About this document

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software. The Data Statement is based on worksheets distributed at the 2020 LREC workshop on Data Statements by Emily M. Bender, Batya Friedman, and Angelina McMillan-Major. This version is adapted from the Markdown template by Leon Dercyznski.

Appendix D

ParlVote+ annotation confusion matrix

Table D.1 (following page) presents a confusion matrix for the policy preference labels provided by the two annotators (Ann1 and Annotator2).
Table D.1: Confusion matrix of policy preference annotations.
Appendix E

ParlVote+ annotation guidelines

E.1 Introduction

This project aims to use codes developed by the Manifesto Project for the annotation of political parties' manifestos (election programmes) to label the policy positions expressed by Members of the United Kingdom Parliament (MPs) in debate motions.

In the House of Commons (the lower chamber of Parliament), each debate normally begins with a motion, a proposal tabled by an MP. These motions express the MPs’ (and usually their parties’) policy positions on one or more topics.

The unit of analysis is the debate motion. For each motion, the aim is to find the most suitable code.

E.2 Policies and coding

The coding scheme is adapted from the Manifesto Project version four (mp v5 https://manifestoproject.wzb.eu/coding_schemes/mp_v5).

There are 57 codes in the scheme, each of which consists of three digits. 56 of these categories are grouped into seven major policy domains. The ‘hundreds’ position in each signifies this category, while the ‘tens’ and ‘units’ positions indicate the specific policy preference.

- Domain 1: External Relations
- Domain 2: Freedom and Democracy
• Domain 3: Political System
• Domain 4: Economy
• Domain 5: Welfare and Quality of Life
• Domain 6: Fabric of Society
• Domain 7: Social Groups

The 57th code is ‘000 No meaningful category applies’, and is used for ‘statements not covered by other categories; sentences devoid of any meaning.’ Further information and examples of coded manifesto excerpts are in the CMP coding Handbook, available at: [https://manifestoproject.wzb.eu/down/papers/handbook_2011_version_4.pdf](https://manifestoproject.wzb.eu/down/papers/handbook_2011_version_4.pdf)

### E.3 Coding notes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Foreign Special Relationships: Positive</td>
<td>Often The USA and Commonwealth countries. Sometimes the Republic of Ireland.</td>
</tr>
<tr>
<td>102</td>
<td>Foreign Special Relationships: Negative</td>
<td>Could be Rep of Ireland.</td>
</tr>
<tr>
<td>103.1</td>
<td>State Centred Anti-Imperialism</td>
<td>Not used in UK Parliament?</td>
</tr>
<tr>
<td>103.2</td>
<td>Foreign Financial Influence</td>
<td>Rare in UK. See 110 European Community/Union: Negative</td>
</tr>
<tr>
<td>104</td>
<td>Military: Positive</td>
<td>Includes support for the Military Covenant, which supports soldiers and veterans. Includes support for military pensions, insurance etc.</td>
</tr>
<tr>
<td>106</td>
<td>Peace</td>
<td>Where terrorism (e.g., Northern Ireland), is discussed in terms of unlawfulness, do not use this code.</td>
</tr>
<tr>
<td>201.2</td>
<td>Human Rights</td>
<td>Use of this code seems broad and extends to e.g. tenants’ and hospital patients’ rights.</td>
</tr>
<tr>
<td>202.1</td>
<td>General: Positive</td>
<td>Generally positive towards democracy rather than negative about lack of democracy (see 304 Political Corruption).</td>
</tr>
<tr>
<td>Code</td>
<td>Topic</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>301</td>
<td>Decentralisation:</td>
<td>Positive covers pro-independence for Scotland, Wales</td>
</tr>
<tr>
<td>302</td>
<td>Centralisation:</td>
<td>Positive covers Unionism</td>
</tr>
<tr>
<td>305.1</td>
<td>Political Authority:</td>
<td>Party Competence: Used to criticise Government or other parties, or to praise own party. Many motions that consist of general criticism of the Government issue fit this code, especially responses to Queen’s speeches, or motions that criticise the Govt. on a range of different policy areas. Also no-confidence motions and calls for General Elections.</td>
</tr>
<tr>
<td>305.2</td>
<td>Political Authority:</td>
<td>Personal Competence: Or a specific Minister</td>
</tr>
<tr>
<td>401</td>
<td>Free Market Economy</td>
<td>Includes many motions in favour of lowering taxes, especially sales taxes.</td>
</tr>
<tr>
<td>402</td>
<td>Incentives</td>
<td>Note that this code applies to incentives to businesses, not consumers. Incentives should generally be something positive proposed rather than criticism of a state of affairs.</td>
</tr>
<tr>
<td>404</td>
<td>Economic Planning</td>
<td>Includes policies aimed at lowering mortgage/interest rate. Consider whether plans or goals (408) are motion’s focus. Should refer to economic planning, not spending plans.</td>
</tr>
<tr>
<td>405</td>
<td>Corporatism/Mixed Economy</td>
<td>Can include immigration policies related to protecting jobs.</td>
</tr>
<tr>
<td>406</td>
<td>Protectionism:</td>
<td>Positive References to balancing the budget, reducing the deficit.</td>
</tr>
<tr>
<td>408</td>
<td>Economic Goals</td>
<td>Consider whether goals or plans (404) are motion’s focus. Should usually not refer to spending plans, which may be a welfare, education etc issue.</td>
</tr>
<tr>
<td>409</td>
<td>Keynesian Demand Management</td>
<td>Positive For this code, the emphasis is on the benefits of spending to the economy.</td>
</tr>
<tr>
<td>415</td>
<td>Marxist Analysis:</td>
<td>Positive Not used in UK debates.</td>
</tr>
<tr>
<td>Code</td>
<td>Description</td>
<td>Guidelines</td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>503</td>
<td>Equality: Positive</td>
<td>May refer to removal of racism, hate speech etc. Can also be economic equality e.g., increasing income tax for high earners, lowering income tax for low earners Could refer to increasing tax on businesses to pay for services. 503 and 504 are quite similar. For 503, the emphasis is usually on an effect on people...</td>
</tr>
<tr>
<td>504</td>
<td>Welfare State Expansion</td>
<td>... whereas 504 is concerned with funding of the welfare system. This code does not include education or infrastructure. Note that public transport, the mail service, and flood defences are infrastructure, not welfare issues. “Welfare” is limited to economic and social issues. Support for Sure Start children’s centres CAN fit this code. Housing should only be included in welfare in the case of public housing, not for general house building to meet demand.</td>
</tr>
<tr>
<td>601.2</td>
<td>National Way of Life: Positive–Immigration: Positive</td>
<td>“New immigrants” can refer to those who are physically in the country, but do not yet have right to remain/residency. 607.1 should be used only for UK residents.</td>
</tr>
<tr>
<td>602.</td>
<td>National Way of Life: Negative–Immigration: Positive</td>
<td>The “new immigrants” may be physically in the country but not yet have residency etc.</td>
</tr>
<tr>
<td>607</td>
<td>Multiculturalism: Positive</td>
<td>Compare 607 and 705: 607 is concerned with benefits to society of immigration, 705 more related to the immigrants themselves.</td>
</tr>
<tr>
<td>701</td>
<td>Labour Groups: Positive</td>
<td>More concerned with defending trade union rights, economic assistance (pay, pensions etc) than 704 Focus here is on people, not institutions like the NHS. Generally regards praise for these groups rather than economic assistance, improved working conditions etc, which might be coded as 701. May include: hospital patients, disabled people, refugees, ethnic minorities, Limited to the UK.</td>
</tr>
<tr>
<td>704</td>
<td>Middle Class and Professional Groups</td>
<td></td>
</tr>
<tr>
<td>705</td>
<td>Underprivileged Minority Groups</td>
<td>May include: hospital patients, disabled people, refugees, ethnic minorities, Limited to the UK.</td>
</tr>
<tr>
<td>706</td>
<td>Non-economic Demographic Groups</td>
<td>May include: young/old people, children. Does not include the unemployed (701). Also, motions concerned with timetabling, parliamentary procedure etc. Finance and budget Bills, which deal with many different financial measures should be coded here, although motions related to individual measures may be coded in Domain 4: Economy.</td>
</tr>
<tr>
<td>000</td>
<td>No meaningful category</td>
<td></td>
</tr>
</tbody>
</table>