# Policy-focused Stance Detection in Parliamentary Debate Speeches

Gavin Abercrombie, Department of Computer Science, University of Manchester
gavin.abercrombie@manchester.ac.uk

Riza Batista-Navarro, Department of Computer Science, University of Manchester
riza.batista@manchester.ac.uk

**Abstract** Legislative debate transcripts provide citizens with information about the activities of their elected representatives, but are difficult for people to process. We propose the task of policy-focused stance detection, in which both the policy proposals under debate and the position of the speakers towards those proposals are identified. We adapt a previously existing dataset to include manual annotations of *policy preferences*, an established schema from political science. We evaluate a range of approaches to the automatic classification of policy preferences and speech sentiment polarity, including transformer-based text representations and a multi-task learning paradigm. We find that it is possible to identify the policies under discussion using features derived from the speeches, and that incorporating motion-dependent debate modelling, previously used to classify speech sentiment, also improves performance in the classification of policy preferences. The proposed use of contextual embeddings and a multi-task learning paradigm do not perform as well as simpler approaches. We analyse the output of the best performing system, finding that discriminating features for the task are highly domain-specific, and that speeches that address policy preferences proposed by members of the same party can be among the most difficult to predict.

## 1 Introduction

Transcripts of legislative debates provide access to information concerning the policies that are publicly supported or opposed by politicians. They are of interest to political scientists, the media, the politicians themselves, and citizens who wish to monitor the activities of their representatives.

However, such documents are complex and difficult for people to process. Transcripts of debates in the United Kingdom (UK) Parliament are so hard for ordinary people to make sense of that parliamentary monitoring website `www.theyworkforyou.com` publishes manually annotated versions of the transcripts. These include crowd-sourced explanations of the debated proposals, as well as policy-focused aggregations of the voting records of parliamentarians. The large quantity and esoteric nature of the data in the parliamentary record (known as *Hansard*) motivates the need for automatic analysis of its contents.

Previous work in the domain of legislative debate transcripts has focused on either *(a)* sentiment polarity classification (Bhavan et al., 2019; Burfoot et al., 2011; Thomas et al., 2006), or *(b)* policy identification (Abercrombie and Batista-Navarro, 2018b; Abercrombie et al., 2019) in isolation. As far as we are aware, these two tasks have not previously been combined in this domain, despite the fact that: *(1)* the information yielded is complementary, and perhaps even necessary, for practical use (i.e., 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.

Borrowing the concept of *policy preferences* from political science, we compare approaches to automatically determining the policy preference that is under discussion in each debate, and whether each speaker supports or opposes it.

**Our contributions** Building on the work of Abercrombie et al. (2019); Abercrombie and Batista-Navarro (2020), we combine policy preference identification and speech-level sentiment polarity analysis to formulate the task of policy-focused speech stance detection for the domain of legislative debate speeches, in which the position of each speaker in a debate is identified in relation to the proposal under discussion. Unlike prior work, we thus obtain interpretable analysis of the positions taken by MPs with respect to the policies presented in parliamentary debates.

To this end, we add a set of manually annotated policy preference labels to a large existing English language corpus of UK parliamentary debates, creating the first dataset to be labelled with both topics (policy preferences) and positions (sentiment) in this domain. We make the enhanced corpus available to the research community.

We use this dataset for the evaluation of approaches to the classification of policy-focused speaker stance. We test classification systems comprising combinations of single- and multi-task learning paradigms, different debate structure models, and varying approaches to text representation and machine learning methods. Our results represent initial benchmarks for this task.

**Research questions** In this paper, we address the following questions:

RQ1 To what extent do humans agree on the policy preference labelling task? We compare agreement between our annotations with those reported in previous work in both political science (Lacewell and Werner, 2013; Mikhaylov et al., 2008) and natural language processing (Abercrombie et al., 2019). The latter found that agreement was comparable for labels applied to debate motions and the manifestos for which the scheme was originally designed, a finding which we re-examine on this new dataset. **Hypothesis H1:** Policy preference labels are as reliable for debate motions as party-political election manifestos.

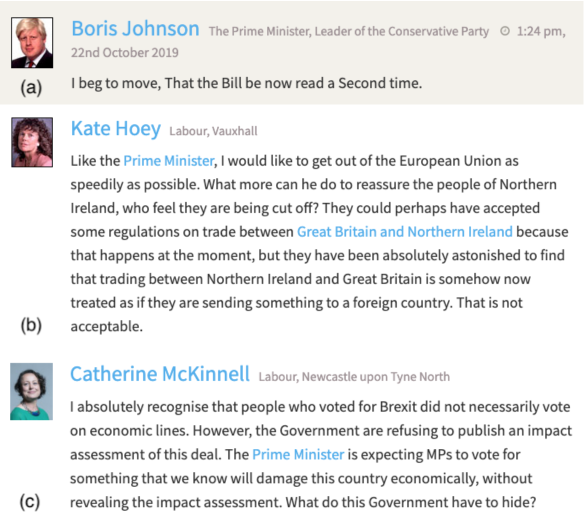
RQ2 How well do machine learning classifiers perform on the combined task of policy-focused stance detection? We test a number of approaches against a majority class baseline. These include fine-tuning pre-trained contextual word embeddings, which we compare to a simple bag-of-words model, and a multi-task learning approach designed to take advantage of mutually beneficial information, which we compare to tackling the constituent tasks independently.
**Hypothesis H2a:** Classification of policy-focused stance will benefit from use of contextual BERT embeddings.
**Hypothesis H2b:** Classification of policy-focused stance will benefit from concurrent classification of policy preferences and speaker sentiment using a multi-task approach.

## 2 Background

**House of Commons debates** As the superior legislative chamber in the UK Parliament, the House of Commons (HoC) draws the attention of the public, the media, and the academic sector, and was therefore chosen as the focus of this study.

Debates in the HoC consist of an opening *motion* (proposal), the content of which usually does not provide clues to the policy that is proposed (see, for example, Figure 1a). We found 75.8 per cent of debate motions in the corpus to contain insufficient information to manually determine a policy preference.

A number of Members of Parliament (MPs) then respond to the motion, when invited to do so by the *Speaker* (the chief presiding officer of the House). An individual MP may make multiple *utterances* during a given debate. Following previous work (Abercrombie and Batista-Navarro, 2020; Salah, 2014; Thomas et al., 2006), we consider a *speech* to be the concatenation of all their utterances in that debate. In many cases, the motion is voted on by MPs in a *division*. As in previous work, we use the record of these votes as labels for sentiment and stance polarity classification.



Boris Johnson The Prime Minister, Leader of the Conservative Party 1:24 pm, 22nd October 2019

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

Kate Hoey Labour, Vauxhall

Like the Prime Minister, I would like to get out of the European Union as speedily as possible. What more can he do to reassure the people of Northern Ireland, who feel they are being cut off? They could perhaps have accepted some regulations on trade between Great Britain and Northern Ireland because that happens at the moment, but they have been absolutely astonished to find that trading between Northern Ireland and Great Britain is somehow now treated as if they are sending something to a foreign country. That is not acceptable.

Catherine McKinnell Labour, Newcastle upon Tyne North

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?

Figure 1: Examples from TheyWorkForYou of *(a)* a debate motion labelled by an annotator with code *110: European Union: Negative*; and two utterances made in response to the motion by speakers who voted *(b) aye* (support) and *(c) no* (oppose).

**Policy preferences** The concept of policy preferences is widely used in political science (Budge et al., 2001) to categorize the positions of politicians. The Manifesto Project (MARPOR: `https://manifestoproject.wzb.eu`) have developed a set of policy preference codes organised under seven '*domains*'. The current coding scheme comprises 74 policy preference codes, almost all of which are 'positional', encoding a positive or negative position towards a policy issue (Mikhaylov et al., 2008). We use these codes as labels for the policy preferences expressed in the debate motions. In the example in Figure 1a, the policy preference label applied to this debate by annotators (see §4.1) is *110: European Union: Negative*.

**Sentiment and stance detection** While use of terminology varies and overlaps in the literature, stance detection can be viewed as a form of sentiment classification. From this perspective, it consists of determining the sentiment polarity of a piece of text towards a predetermined 'given target of interest' (Mohammad et al., 2016). In the case of parliamentary debates, for each example speech, we seek to determine *(1)* the nature of its target—the policy preference under debate—and *(2)* the position or sentiment expressed towards it—*support* or *opposition*. We consider the combined policy preference and speech sentiment labels to represent the speaker's stance on a particular policy. For instance, in the example in Figure 1, the stance of speech extracts *(b)* and *(c)* are *European Union: Negative—support* and *European Union: Negative—oppose*, respectively.

## 3 Related work

Sentiment classification is one of the the most active areas of research in natural language processing. Within the domain of legislative debates, examples include classification of speeches from the US Congress (Burfoot et al., 2011; Ji and Smith, 2017; Proksch et al., 2019; Thomas et al., 2006), and the UK Parliament (Abercrombie and Batista-Navarro, 2018b, 2020; Bhavan et al., 2019; Salah, 2014; Sawhney et al., 2020). In these works—and in common with ours—speaker sentiment is assumed to be analogous to vote outcome. However, in the task undertaken in these previous works, the nature of the targets—the Bills or motions under debate—is not identified.

The related task of stance detection—in which the target of sentiment *is* (pre-)determined—has been applied to such domains as social media (e.g. Augenstein et al., 2016a,b; Hardalov et al., 2021; Li et al., 2021; Mohammad et al., 2016), online debate forums (e.g. Hardalov et al., 2021; Hasan and Ng, 2013; Somasundaran and Wiebe, 2010; Sridhar et al., 2015), and news articles (Ferreira and Vlachos, 2016; Schiller et al., 2021). For a recent survey, see Küçük and Can (2020).

In most of this work the target is pre-chosen by the user or the system. In the political domain, this has been framed as agreement detection in which two pieces of text are compared (Menini and Tonelli, 2016; Menini et al., 2017), or classification of *support* or *attack* towards pre-defined policies (Menini et al., 2018). While Vamvas and Sennrich (2020) carry out stance detection on the positions expressed by Swiss politicians, they do not perform automatic identification of the policies discussed, only conducting binary *in favour/against* classification in a similar vein to the sentiment/position classification work discussed above.

More similarly to this work, Bar-Haim et al. (2017) used a supervised approach to identify both the stances of extracts from Wikipedia articles and the targets of those stances from a closed list of 'controversial topics'. However, this labelling scheme does not cover the policy positions proposed in parliament.

A common framework for stance detection is the SDQC (Support-Deny-Query-Comment) annotation scheme of Zubiaga et al. (2016). While potentially suitable for our data (*support* and *deny* are equivalent to our *support* and *oppose* labels), application of this framework would require manual annoation of each instance in the dataset with the more fine-grained labels. Instead, we follow the majority of work on legislative debates (e.g. Abercrombie and Batista-Navarro, 2018a; Thomas et al., 2006; Salah, 2014) in taking advantage of pre-existing vote-derived binary labels at the speech level, and thus only requiring the addition of policy preference labels for each debate.

In most of the reviewed work, stance targets are explicitly selected by the authors of the task (e.g. *Donald Trump* (Augenstein et al., 2016a,b), *Richard Nixon* and *John F. Kennedy* (Menini et al., 2018), or *atheism* (Mohammad et al., 2016)). Unlike these, we frame target selection as a multiclass topic classification problem, making use of an existing schema validated by political scientists.

Document classification is an active area of research for tasks such as identification of news and Wikipedia categories (Zhang et al., 2015). For classification of HoC debates, Abercrombie and Batista-Navarro (2018b) used 'policy' labels crowdsourced by the parliamentary monitoring website `https://www.publicwhip.org.uk/` but found this framework limited as it could not be easily scaled up from the small existing labelled dataset. Abercrombie et al. (2019) created a manually annotated dataset of policy preferences in debate motions, and achieved promising results in classifying debate motions according to the MARPOR coding scheme. However, this corpus is unsuitable for our purposes as: *(1)* it does not include speeches made in response to the motions; and *(2)* the motions in this 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, we seek a mixture of motion types, more represenative of the Hansard record as a whole. Additionally, while they classified debate motions with policy preference labels using textual features derived from the motions themselves, many of the motions in Hansard—and in the corpus used in this study—contain little in the way of informative textual content (Figure 1a is a typical example). Rather than the motions, we therefore rely on features derived from the response speeches, which we use as input for the classification of both motions and speeches.

Multi-task learning approaches have been taken to many tasks, including part-of-speech tagging, chunking, and named entity recognition (Collobert and Weston, 2008). While such approaches have been applied to sentiment classification of customer reviews (Yu and Jiang, 2016), we are not aware of any uses of multi-task learning in the legislative debate domain. The most common approach to multi-task learning—which we compare with the single task paradigm—is that of hard parameter sharing, first proposed by Caruana (1993).

# 4 Data

*ParlVote* (Abercrombie and Batista-Navarro, 2020) is a large corpus (34,010 examples) of HoC debate speeches made between from 1997 and 2019. Each example speech consists of the concatenated utterances of an individual speaker in a given debate, and is presented with the debate motion to which it responds, as well as the vote of the speaker (in support or opposition to the motion), and metadata associated with the debate and the speakers. We adapted this corpus to include an additional, manually annotated policy preference label for each example. As capitalization can be informative in this domain (for example in the terms of address '*Friend*', '*Lady*', '*Gentleman*'), we did not lowercase the text.

## 4.1 Annotation

We adapted the ParlVote annotation guidelines to include the new codes used in the updated MARPOR *Coding Scheme version 5* (Werner et al., 2015). We make our guidelines available at `https://tinyurl.com/y5twunrm`.

The first author of this paper annotated each debate motion following these guidelines. Included in the guidelines were 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 MPs may use such motions politically, on the face of it they are concerned simply with the running of Parliament, rather than policy.

- Debates with divisions that are not on the motion in question. In many cases the division held at the end of the debate is held on some other point that has been brought up during the debate, such as an amendment introduced by the Speaker.

- Motions that appear to fit several codes, such as *Finance* Bills, *Local Finance* Bills, and Bills concerning the budgets of e.g., Police forces. Within the area of budgetary Bills is the exception of motions debates concerning approval of European Union (EU) Finance Bills, which tend to be positive or negative about the EU.

- Motions concerning constituency boundary changes.

We excluded all examples given this label from the dataset used for the experiments reported below as they cover a wide range of topics and/or do not fit into any of the Manifesto Project codes. While 56 of the policy preference codes were used as labels by the annotators, we also excluded all examples with policy preference codes that occur fewer than 100 times in the dataset, leaving 34 codes used in the classification experiments. This left 23,181 example speeches given by 1,321 unique MPs given in response to 1,215 different debates. Each example has a manually annotated policy preference label and a vote-derived speech stance polarity label. Of these, 305.1: *Political Authority: Party Competence* is the most common, with 4,926 labelled examples (see Appendix A).

Each instance in the corpus also retains it's *support/oppose* label from the original ParlVote corpus, which we use to label the stance taken in each speech towards the policy under debate.

## 4.2 Inter-annotator agreement

In order to validate the new motion policy preference labels, we recruited a second annotator to label a randomly selected subsection of the corpus. After annotation, comparison, and discussion of some initial training examples, she labelled 108 motions (8.9% of the total). On this subset, we calculated a Cohen's *kappa* agreement score of 0.38, which can be interpreted as representing '*fair*' (Landis and Koch, 1977) or '*poor*' (Fleiss et al., 1981) agreement. This is comparable to other studies of annotation using the Manifesto Project codes (Lacewell and Werner, 2013; Mikhaylov et al., 2008), and similar to agreement on election manifestos for which the labelling scheme was originally designed (Abercrombie et al., 2019). The level of agreement highlights that this is a non-trivial task on which agreement between different human annotators is difficult to achieve. Despite this issue of annotation reproducibility, these labels are considered to be valid by political scientists—as evidenced by Volkens et al. (2015), who found 230 articles that use this annotated data in the eight journals they examined. With comparable inter-annotator agreement, we consider them to be the best available labelling scheme for our task.

We make the adapted dataset, *ParlVote+*, available for the research community at: `https://tinyurl.com/y22rrta7`.[1] There, we also provide a full data statement, following the guidelines of Bender and Friedman (2018).

## 5 Method

We investigate approaches to determining, for each example in the dataset, *(a)* the policy preference expressed in the debate motion, and *(b)* the sentiment (position) expressed in the speech towards that motion: *support* (positive) or *oppose* (negative).

We compare the performance of systems comprised of combinations of the following:

- Learning paradigms (see Figure 2):
  - Single tasks: inputs are processed separately for the two tasks, as in previous work.
  - Multi-task learning: we 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: Abercrombie and Batista-Navarro (2018a) showed that Government-proposed motions tend to be positive and those tabled by opposing parties negative, and that this could be used as a proxy for the polarity of the motions. We classify examples from debates initiated by members of the governing and opposition parties separately.
- Text representations:
  - Bag-of-words (BOW): we used term frequency-inverse document frequency (tf-idf) scores of terms in the dataset to select unigram features (as previous work suggests that the addition of higher *n*-gram features does not improve perfomance in this domain (Abercrombie and Batista-Navarro, 2018a)). Aside from not lowercasing the text, we used the default settings from scikit-learn to tokenize and extract ti-idf features from the texts.[2]
  - Contextual word embeddings: we fine-tuned BERT embeddings (Devlin et al., 2019) on our classification tasks. Systems using this approach have achieved state-of-the-art performances, and have been applied to the two tasks of interest in this domain (Abercrombie et al., 2019; Abercrombie and Batista-Navarro, 2020). As we included uppercase characters in the input, we used the *large, cased* version, available at `https://tfhub.dev/google/bert_cased_L-12_H-768_A-12/`. We use Google's BERT tokenizer,[3] and pad the texts to the maximum input of 512 tokens, then fine-tune the top 3 layers of the BERT model. The (fine-tuned) final layer of BERT embeddings is then used as input to one of the following neural classifiers.
- Machine learning classification algorithms. We used neural networks of two hidden layers, with the second of these separated into two task-specific layers in the multi-task learning setting (see Figure 2). We used Adam optimization with a learning rate of $1 * 10^{-5}$, a batch size of 32 and, with the BOW input only, a dropout rate of 0.5 for each layer.[4] For binary (speech sentiment) and multiclass (motion policy preference), we used sigmoid and softmax activation layers,

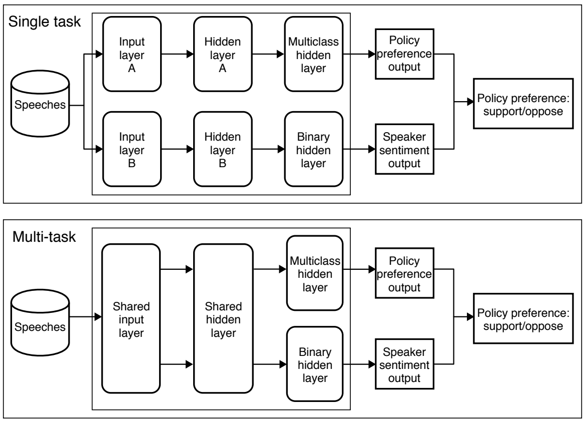


Figure 2: Single and multi-task learning paradigms.

[1] Note this URL links to an anonymised Google Drive folder. Link to a permanent data repository will be provided on acceptance.

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

[3] `https://github.com/google-research/bert/blob/master/tokenization.py`

[4] Opitimzation experiments showed that dropout negatively influenced the performance using BERT (see Appendix B).

| Learning | Text | Machine learning | Policy pref. | | Sentiment | | Policy-focused stance | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| paradigm | representation | method | Ind. | Dep. | Ind. | Dep. | Mean | | Absolute | |
| — | — | Majority class | 1.1 | 1.1 | 35.8 | 35.8 | 18.5 | 18.5 | 0.3 | 0.3 |
| Single-task | BOW | MLP | 58.0 | **64.1** | 61.2 | 70.8 | 59.6 | **67.4** | 33.3 | **45.2** |
| | | CNN | 53.1 | 59.5 | 61.5 | 70.1 | 57.3 | 64.8 | 29.9 | 40.8 |
| | BERT | MLP | 50.4 | 57.2 | 61.1 | 67.6 | 55.8 | 62.4 | 28.7 | 36.4 |
| | | CNN | 43.0 | 52.5 | 64.3 | 71.7 | 53.7 | 62.1 | 25.2 | 35.6 |
| Multi-task | BOW | MLP | 56.0 | 52.7 | 63.9 | **74.3** | 60.0 | 63.5 | 34.1 | 38.2 |
| | | CNN | 38.2 | 38.5 | 58.5 | 68.8 | 48.4 | 53.7 | 19.9 | 21.8 |
| | BERT | MLP | 50.9 | 43.7 | 60.1 | 72.8 | 55.5 | 58.2 | 27.9 | 29.1 |
| | | CNN | 44.4 | 41.1 | 59.4 | 70.6 | 51.9 | 55.8 | 23.9 | 25.4 |

Table 1: Macro-averaged 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.

respectively. We used early stopping and tested on the model that performed best on the validation set. Hyperparameters were chosen based on optimisation experiments, the results of which are presented in Appendix B.

We compared the following classes of network:

- Multi-layer perceptron (MLP): we used a network with hidden layers of 512 nodes and ReLU activation.
- Convolutional neural network (CNN): a network of one-dimensional convolutional layers with 512 filters, convolution windows spanning three tokens, and max pooling.

We used a randomly sampled 80/10/10 split of the data. The experiments can be reproduced using our python notebook, which we make available with all code and data at `https://tinyurl.com/y62jrkyt`.

## 6 Results

We evaluated the systems described above against the majority class for each task. Slight differences in these baseline scores in the motion-dependent and independent settings arise from variations in the class distributions in the test sets in these settings. Due to the class imbalances in the dataset, we report the macro-weighted F1 score as the evaluation metric.

### 6.1 Overall results

Results are presented in Table 1. Here, *policy-focused stance* represents the *sentiment polarity* of speakers towards the *policy preference* under debate. We report two measures of this for each system configuration: *(1)* the mean of the F1 scores for policy preference identification and sentiment classification, and *(2)* the absolute F1 where only examples for which both predicted labels match the true class labels are considered to be correct.

Most of the tested system configurations outperformed the naive baselines. In most cases, the motion dependent models performed better than those that did not take into account this aspect of debate structure. Overall, contrary to our hypotheses, neither BERT nor the multi-task learning paradigm improved performance over the BOW and single-task set-ups. 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 the with loss of information due to BERT's maximum sequence length. The MLP classifier performed better than the CNN in nearly all scenarios. The highest overall F1 score for the combined tasks (67.4 mean, 45.2 absolute) 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 by Abercrombie et al. (2019), despite using completely different input texts, having no access to the content of the motions themselves.

### 6.2 Results using shorter input speeches

The lower, poorer performance of BERT text representations in all settings is perhaps due to its the 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, we also ran the single task MLP classifier on a subset of the data consisting solely of the 13, 162 speeches in the dataset that consist of 512 tokens or fewer (calculated using the scikit-learn tokenizer). Results of these experiments are shown in Table 2.

F1 scores here are lower than when using the full

| Text | Policy | | Speech | | Policy-focused stance | | | |
|---|---|---|---|---|---|---|---|---|
| representation | preference | | sentiment | | Mean | | Absolute | |
| | Ind. | Dep. | Ind. | Dep. | Ind. | Dep. | Ind. | Dep. |
| Majority class | 0.1 | 0.1 | 36.0 | 36.0 | 18.1 | 18.1 | 0.3 | 0.3 |
| BOW | 32.6 | 40.9 | **56.3** | 58.3 | **44.5** | 49.6 | 17.7 | 19.3 |
| BERT | **34.9** | **45.0** | 51.0 | **62.8** | 43.0 | **53.9** | **18.5** | **24.8** |

Table 2: Macro-averaged F1 scores for classification of policy preference (multiclass), speech sentiment (binary) and policy-focused stance (mean of these scores) using BOW and BERT-based text representations in the *single-task–MLP* classification setting on shorter speeches of 512 tokens or fewer.

| Code | Policy pref. | Sentiment | Stance (mean) | Code | Policy pref. | Sentiment | Stance (mean) |
|---|---|---|---|---|---|---|---|
| *104* | 83.8 | 68.4 | 76.1 | *411* | **84.6** | **81.2** | **82.9** |
| *105* | 57.1 | 47.5 | 52.3 | *413* | 45.5 | 72.7 | 59.1 |
| *106* | 76.2 | 61.1 | 68.7 | *501* | 65.0 | 46.4 | 55.7 |
| *108* | 67.5 | 58.6 | 63.1 | *503* | 46.7 | 69.3 | 58.0 |
| *110* | 65.9 | 54.9 | 60.4 | *504* | 65.4 | 75.0 | 70.2 |
| *201.2* | 56.9 | 55.3 | 56.1 | *505* | 78.2 | 67.3 | 72.8 |
| *202.4* | 76.9 | 76.4 | 76.7 | *506* | 50.0 | 74.7 | 62.4 |
| *203* | 31.6 | 57.8 | 44.7 | *507* | 56.1 | 69.3 | 62.7 |
| *204* | 69.8 | 55.2 | 62.5 | *601.2* | 36.4 | 59.0 | 47.7 |
| *301* | 54.5 | 67.0 | 60.8 | *602.2* | 36.4 | 47.6 | 42.0 |
| *302* | 41.0 | 54.5 | 47.8 | *603* | 52.8 | 60.0 | 56.4 |
| *304* | 52.6 | 47.4 | 50.0 | *604* | 69.8 | 53.7 | 61.8 |
| *305.1* | 83.5 | 74.6 | 79.1 | *605.1* | 79.4 | 66.4 | 72.9 |
| *305.2* | 33.3 | 59.0 | 46.2 | *605.2* | 60.8 | 64.7 | 62.8 |
| *401* | 51.8 | 68.3 | 60.1 | *701* | 48.5 | 71.8 | 60.2 |
| *402* | 44.7 | 64.4 | 54.6 | *702* | 47.1 | 71.7 | 59.4 |
| *403* | 61.1 | 62.0 | 61.6 | *706* | 42.1 | 78.9 | 60.1 |

Table 3: F1 scores for policy preference, sentiment, and (mean) policy-focused stance by policy preference code. Highest scores for each task are bold, contrastive pairs of policy preference codes in grey boxes.

dataset due to the smaller size of the training set. However, the fact that under these conditions use of BERT outperforms BOW, shows the importance of providing BERT with the full speech, and indicates that where this is possible fine-tuning on BERT should lead to improved performance over the BOW model.

## 6.3 Results by policy preference class

Examining the performance of one of the best performing system configurations—the *single-task–BOW–MLP–motion-dependent* system—for each (*true*) policy preference label (Table 3), there are a wide variety of scores for each task.

Each policy preference class received between four and 21 predicted labels in the classifier output ($\mu$ = 10.4). Labels with contrastive pairs did not necessarily seem to be more difficult to predict than individual class labels, with, for example *104: Military: Positive* obtaining one of the highest F1 scores for policy preference detection. Similarly, 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 the highest scores overall. This suggests that the model can struggle to differentiate between the closely related, but opposing policy preference classes.

264 examples (22.1% of errors) were classified incorrectly for both policy preference and stance, 520 (43.6%) for policy preference only, and 410 (34.3%) for stance only. Figure 3 shows the predicted policy preference labels with respect to the true labels assigned by the annotators. Where mis-classifications occur, the classifier does not tend to prefer closely related labels, with more than double the number of out-of-*domain* (69.9%) to in-*domain* (31.1%) mis-classifcations. This suggests considerable overlap of language use in policy *domains* such as *4: Economy* and *5: Welfare and Quality of Life*, where issues relating to both may frequently be discussed in the same debates, and on which the annotators frequently disagreed.

## 6.4 System output analysis

To gain an understanding of the challenges involved in improving classification performance on these tasks, we examined in closer detail the output of the *single-*

| | All | + | - | Gov. | Opp. | Own | Other | Gov.+ | Gov.- | Opp.+ | Opp.- |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Max** | 0.44 | 0.44 | 0.28 | 0.25 | 0.44 | 0.38 | 0.44 | 0.25 | 0.23 | 0.44 | 0.38 |
| **Mean** | -0.01 | -0.01 | -0.02 | -0.02 | -0.01 | -0.01 | -0.02 | -0.02 | -0.02 | -0.01 | -0.01 |
| **Min** | -0.38 | -0.38 | -0.38 | -0.38 | -0.38 | -0.38 | -0.38 | -0.38 | -0.31 | -0.38 | -0.31 |

| Own | Own | Oth. | Oth. | Gov. | Gov. | Gov | Gov- | Opp. | Opp. | Opp. | Opp. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| + | - | + | - | own+ | own- | oth+ | oth- | own+ | own- | oth+ | oth- |
| 0.25 | 0.38 | 0.25 | 0.25 | 0.25 | 0.21 | 0.19 | 0.25 | 0.25 | 0.38 | 0.44 | 0.28 |
| -0.01 | -0.02 | -0.01 | -0.02 | -0.02 | 0.01 | -0.02 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 |
| -0.38 | -0.31 | -0.38 | -0.31 | -0.31 | -0.381 | -0.31 | -0.38 | -0.38 | -0.31 | -0.38 | -0.31 |

Table 4: 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.

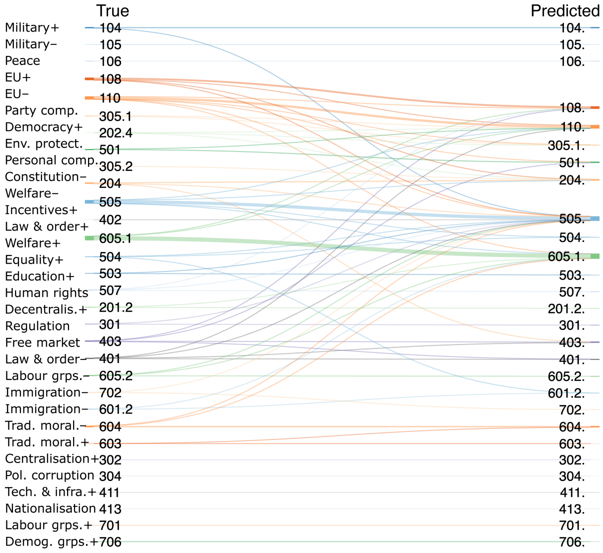


Figure 3: True policy preference labels and the labels predicted by the classifier.

*task–BOW–MLP–motion-dependent* system.

### 6.4.1 Features of speech polarity

In these experiments, we found that performance was improved by modelling debate structure in the motion-dependent setting. This supports the findings of Abercrombie and Batista-Navarro (2018a), who observed that the textual features that discriminated between supportive and oppositional speeches were not typically positive or negative when used in other domains.

To investigate how sentiment is manifested in this domain, we 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 4.

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 expecteed to do in general usage.

To examine which terms in the speeches *do* indicate sentiment, we obtained the permutation importance scores of each unigram in the input vocabulary. That is, for feature $j$ in the feature set $N$, we calculated the permuation feature importance as the difference between performance (in this case, the F1 score) using the original datset $D$ and a corrupted version $\tilde{D}$, in which $j$ has been randomly shuffled (Breiman, 2001). We consider features with higher scores to be more important to the model. A sample of the most important

| Motion-independent | | Motion-dependent | | | |
|---|---|---|---|---|---|
| All | | Government | | Opposition | |
| *Labour* | +0.13 | *approach* | +0.17 | *Minister* | 0.00 |
| *Gentleman* | +0.13 | *average* | 0.00 | *Opposition* | −0.07 |
| *shadow* | −0.09 | *costs* | −0.03 | *Prime* | +0.09 |
| *Prime* | +0.09 | *police* | +0.13 | *welcome* | +0.19 |
| *party* | 0.00 | *contrast* | 0.00 | *shadow* | −0.09 |
| *cuts* | +0.01 | *registration* | 0.00 | *Is* | +0.02 |
| *Lady* | 0.00 | *officers* | 0.00 | *continue* | 0.00 |
| *situation* | −0.08 | *proposals* | 0.00 | *best* | +0.38 |
| *threat* | −0.28 | *hit* | −0.03 | *look* | 0.00 |
| *outside* | 0.00 | *tier* | +0.06 | *Members* | 0.00 |
| *pay* | +0.06 | *fees* | +0.19 | *Secretary* | 0.00 |
| *Lords* | 0.00 | *chance* | +0.08 | *ensure* | 0.00 |
| *crisis* | −0.06 | *labour* | +0.13 | *way* | +0.01 |
| *Government* | 0.00 | *constituency* | 0.00 | *suggestion* | −0.05 |
| *constituents* | 0.00 | *dealt* | 0.00 | *public* | −0.04 |
| *wants* | −0.06 | *running* | 0.00 | *motion* | 0.00 |
| *important* | +0.08 | *data* | 0.00 | *Clearly* | +0.19 |
| *careful* | +0.19 | *willingness* | +0.13 | *support* | +0.09 |
| *week* | 0.00 | *tackling* | 0.00 | *worse* | −0.29 |
| *stop* | −0.02 | *strategy* | +0.06 | *said* | 0.00 |

Table 5: 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.

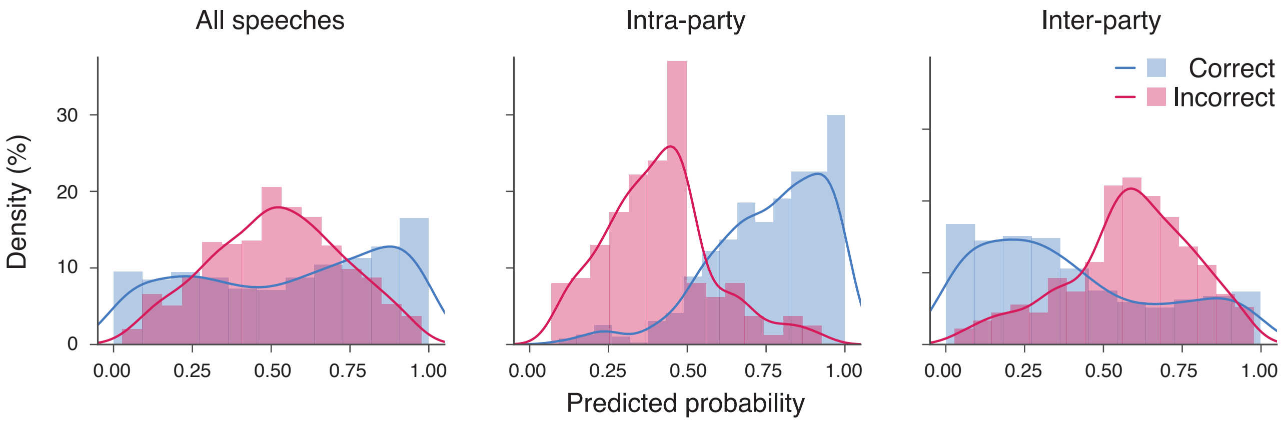


Figure 4: Distribution frequencies (histograms and density curves) of the correct and incorrect predicted probabilities of sentiment labels being positive for three categories of speech: those by *all* speakers, and *intra-* and *inter-party* responses.

features (the top 20) in each setting according to this metric is shown in Table 5

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.

### 6.4.2 Party affiliations

As MPs usually vote along party lines, it would be possible to achieve good sentiment classification results by setting a classifier to make predictions on that simple basis alone. 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 (Proksch and Slapin, 2015). To investigate how this effects sentiment polarity classification, we compared the performance of *rebel* MPs—those voting against a motion proposed by their own party or in support of one proposed by an-

|  | Stance ✓ | PP ✓, sent. ✗ | Sent. ✓, PP ✗ | Stance ✗ |
|---|---|---|---|---|
| *n* examples | 1120 | 404 | 492 | 303 |
| Max. tokens | 20730 | 6505 | 6484 | 4742 |
| Mean tokens | 876.9 | 916.5 | 761.4 | 867.6 |
| Min. tokens | 2 | 2 | 2 | 2 |
| Std. deviation | 1213.9 | 953.2 | 1115.6 | 4742 |
| < 50 tokens | 103 | 43 | 61 | 35 |
| >= 50 tokens | 1017 | 361 | 431 | 268 |

Table 6: Number of speeches by token counts and prediction outcome (✓= correct and ✗= incorrect).

other party—and *loyal* MPs. This produced F1 scores of 77% and 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 4 present a closer look at this. They show the predicted probabilites of examples being assigned to the positive class. We 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 probablities that it assigns to the positive class (correctly, more often than not). Meanwhile, negative predictions (usually incorrect) are made with probabilites 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 about the nuances of individual speaker stance.

### 6.4.3 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 6).

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 (Abercrombie and Batista-Navarro, 2018a; Salah, 2014). There are 2,941 such speeches in ParlVote, which are fairly balanced between the positive and negative classes (53/47%) and a very similar distibution of policy preference labels as the main dataset. 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 hear*', '*Under Labour*'–both *negative* stance) 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 Discussion and conclusion

Policy-focused stance detection of parliamentary speeches is a challenging task, which we have framed as combined binary and multiclass classification. For this, we enhanced an existing dataset with an additional set of policy preference labels. While inter-annotator agreement on policy preference labels is modest, it is similar to that reported in previous work on both parliamentary debates and election manifestos. To address the issue of low annotator agreement, and the fact that classifiers frequently misclassify speeches across policy *domains*, future work could take a *perspectivist* approach to annotator disagreement (Basile et al., 2021a,b), and consider reframing the task as a multiclass *and multilabel* problem, in which more than one policy preference code may be valid per speech. Notwithstanding this issue, and despite the large number of classes in the policy prediction task, and the fact that the input features we used were based only on the content of speeches (not the motions or titles, as in previous work (Abercrombie et al., 2019)), we have been able to obtain reasonable results, comfortably beating the majority class baselines.

Modelling of the structure of parliamentary debates in the form of motion-dependent classification was seen to improve performance on speech sentiment classification in prior work. In this study, we found that it is not only consistently superior for speech sentiment classification, but also improves the identification of policy preferences, the topics under discussion. We have shown that the differences between supportive and opposing speeches do not derive from generally sentiment bearing words, 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, we 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 plenty of scope for further experimentation with more complex architectures for this approach.

In these experiments, fine-tuning on BERT embeddings led to considerably worse performances. Considering the widespread successes of this approach, this also warrants further investigation. With recent work suggesting that, for real-world tasks and datasets, pretraining the embeddings on in-domain data may be necessary (Xia et al., 2020), a more domain-specific approach may be desirable.

While other work on sentiment and stance detection in the domain of parliamentary debates has effectively overlooked the targets of those opinions, we have combined approaches to sentiment and topic detection to formulate a task with potential for real-world application. Although there remains much room for improvement in classification performance, we have shown that the task of policy-focused speech stance detection can be feasibly automated, even with simple features and neural architectures. Although we have focussed our annotation effort and analysis on debates from the UK Parliament, the proposed approach is generalisable to other legislatures, or indeed any political debates that feature proposed motions and supporting and opposing documents.

In future work, we will focus on refining the annotation scheme in order to obtain greater labelling consistency and improved classification performance, as well as adapting the methods for the legislative debate domain.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions and Nancy Greig for her diligent work on data annotation.

## References

Abercrombie, Gavin and Riza Batista-Navarro. 2018a. 'aye' or 'no'? speech-level sentiment analysis of hansard UK parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abercrombie, Gavin and Riza Batista-Navarro. 2020. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Abercrombie, Gavin and Riza Theresa Batista-Navarro. 2018b. Identifying opinion-topics and polarity of parliamentary debate motions. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 280–285, Brussels, Belgium. Association for Computational Linguistics.

Abercrombie, Gavin, Federico Nanni, Riza Batista-Navarro, and Simone Paolo Ponzetto. 2019. Policy preference detection in parliamentary debate motions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 249–259, Hong Kong, China. Association for Computational Linguistics.

Augenstein, Isabelle, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016a. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Augenstein, Isabelle, Andreas Vlachos, and Kalina Bontcheva. 2016b. USFD at SemEval-2016 task 6: Any-target stance detection on Twitter with autoencoders. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 389–393, San Diego, California. Association for Computational Linguistics.

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

Bar-Haim, Roy, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Basile, Valerio, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021a. Toward a Perspectivist turn in ground truthing for predictive computing. In *Conference of the Italian Chapter of the Association for Intelligent Systems (ItAIS 2021)*.

Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021b. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Bender, Emily M. and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bhavan, Anjali, Rohan Mishra, Pradyumna Prakhar Sinha, Ramit Sawhney, and Rajiv Ratn Shah. 2019. Investigating political herd mentality: A community sentiment based approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 281–287, Florence, Italy. Association for Computational Linguistics.

Breiman, Leo. 2001. Random forests. *Machine Learning*, 45:5–32.

Budge, Ian, Hans-Dieter Klingemann, et al. 2001. *Mapping policy preferences: Estimates for parties, electors, and governments, 1945-1998*, volume 1. Oxford University Press.

Burfoot, Clinton, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1506–1515, Portland, Oregon, USA. Association for Computational Linguistics.

Caruana, Richard A. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning Proceedings 1993*, pages 41 – 48. Morgan Kaufmann, San Francisco (CA).

Collobert, Ronan and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

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

Ferreira, William and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Fleiss, Joseph L, Bruce Levin, and Myunghee Cho Paik. 1981. *Statistical methods for rates and proportions*. John Wiley & sons.

Hardalov, Momchil, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hasan, Kazi Saidul and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

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

Küçük, Dilek and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Lacewell, Onawa P and Annika Werner. 2013. Coder training: key to enhancing reliability and validity. *Mapping Policy Preferences from Texts*, 3:169–194.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Li, Yingjie, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Menini, Stefano, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Menini, Stefano, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944, Copenhagen, Denmark. Association for Computational Linguistics.

Menini, Stefano and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2461–2470, Osaka, Japan. The COLING 2016 Organizing Committee.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2008. Coder reliability and misclassification in Comparative Manifesto Project codings. In *the 66th MPSA Annual National Conference*.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Proksch, Sven-Oliver and Jonathan B Slapin. 2015. *The politics of parliamentary debate.* Cambridge University Press.

Rogers, Robert and Rhodri Walters. 2015. *How Parliament works.* Routledge.

Ruder, Sebastian. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

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

Sawhney, Ramit, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2020. GPolS: A contextual graph-based language model for analyzing parliamentary debates and political cohesion. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4847–4859, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Schiller, Benjamin, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *Künstliche Intelligenz.*

Somasundaran, Swapna and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Sridhar, Dhanya, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 116–125, Beijing, China. Association for Computational Linguistics.

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Vamvas, Jannis and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection.

Volkens, Andrea, Cristina Ares, Radostina Bratanova, and Lea Kaftan. 2015. Scope, range, and extent of Manifesto Project data usage: A survey of publications in eight high-impact journals. In *Handbook for Data Users and Coders.* WZB.

Werner, Annika, Onawa Lacewell, and Andrea Volkens. 2015. Manifesto coding instructions: 5th fully revised edition.

Xia, Patrick, Shijie Wu, and Benjamin Van Durme. 2020. Which *BERT? A survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online. Association for Computational Linguistics.

Yu, Jianfei and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246, Austin, Texas. Association for Computational Linguistics.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.

Zubiaga, Arkaitz, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.

# A The ParlVote+ corpus

Table 7 shows the number of example speeches that are labelled with each of the MARPOR codes.

# B Machine learning parameter optimisation results

Results of preliminary experiments to select the optimal size of CNN window, number of layers of BERT to fine-tune, and dropout rate are shown in Tables 8, 9, and 10.

For the main experiments, the results of which are presented in Section 6, we selected the parameters that resulted in highest F1 scores in the majority of settings in these preliminary tests: CNN window size of 3, fine-tuning three layers of BERT, and a dropout rate of 0.5 in the BOW setting, with no dropout when using BERT.

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

Table 7: 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 5 are highlighted in bold text.

| Window size | Text representation | Learning paradigm | Policy pref. | | Sentiment | | Policy-focused stance | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ind. | Dep. | Ind. | Dep. | Mean | | Absolute | |
| 3 | BOW | STL | **53.1** | **59.1** | **61.5** | **70.1** | **57.3** | **64.8** | **30.0** | **40.8** |
| | | MTL | **38.2** | **38.5** | **58.5** | **68.8** | **48.4** | **53.7** | **19.9** | **21.8** |
| | BERT | STL | **43.5** | **51.3** | **64.0** | **70.1** | **53.8** | **60.1** | **26.3** | **33.9** |
| | | MTL | 38.3 | **42.8** | 56.3 | 71.0 | 47.3 | **56.9** | 18.0 | **27.9** |
| 4 | BOW | STL | 32.6 | 40.4 | 56.3 | 58.2 | 44.5 | 49.3 | 17.7 | 19.3 |
| | | MTL | 1.0 | 1.45 | 36.0 | 37.6 | 18.5 | 19.5 | 0.3 | 0.5 |
| | BERT | STL | 21.5 | 30.6 | 54.7 | 64.0 | 38.1 | 47.3 | 11.4 | 16.4 |
| | | MTL | 36.7 | 25.4 | 57.2 | 71.8 | 46.9 | 48.6 | 18.0 | 16.2 |
| 5 | BOW | STL | 52.5 | 41.4 | 61.0 | 66.2 | 56.8 | 53.8 | 29.6 | 23.9 |
| | | MTL | 0.40 | – | 36.0 | – | 18.2 | – | 0.1 | – |
| | BERT | STL | 21.8 | 26.9 | 50.9 | 51.1 | 36.3 | 39.0 | 10.5 | 12.1 |
| | | MTL | **39.4** | 29.7 | **57.8** | **72.3** | **48.6** | 51.0 | **21.5** | 20.3 |

Table 8: Macro F1 scores for classification using CNN with windows of three, four, and five tokens.

| Fine-tune layers | Learning paradigm | Policy pref. | | Sentiment | | Policy-focused stance | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ind. | Dep. | Ind. | Dep. | Mean | | Absolute | |
| 3 | STL | **50.4** | **57.2** | **61.2** | 67.6 | **55.8** | 62.4 | **28.7** | **36.4** |
| | MTL | **50.9** | 43.7 | 60.1 | **72.8** | **55.5** | **58.2** | **27.9** | **29.1** |
| 6 | STL | 45.7 | 53.6 | 61.1 | **73.0** | 53.4 | **63.3** | 24.9 | **37.1** |
| | MTL | 36.0 | 36.2 | **64.7** | 63.9 | 50.4 | 50.1 | 21.0 | 21.4 |
| 9 | STL | 41.1 | 54.0 | 59.7 | 70.8 | 50.4 | 62.4 | 22.5 | 35.7 |
| | MTL | 37.7 | **45.2** | 63.4 | 60.2 | 50.5 | 52.7 | 22.8 | 24.7 |

Table 9: Macro F1 scores for classification using MLP and fine-tuning three, six, and nine of the 12 BERT layers. Highest F1 scores for each learning paradigm are presented in bold, absolute highest scores are underlined.

| Dropout rate | Text representation | Learning paradigm | Policy pref. | | Sentiment | | Policy-focused stance | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ind. | Dep. | Ind. | Dep. | Mean | | Absolute | |
| 0.5 | BOW | STL | **58.0** | **64.1** | **61.2** | **70.8** | **59.6** | **67.4** | **33.3** | **45.2** |
| | | MTL | **56.0** | **52.7** | 63.9 | **74.2** | **60.0** | **63.4** | **34.1** | 38.2 |
| | BERT | STL | 47.5 | 53.2 | 60.0 | **70.6** | 53.7 | 61.9 | 24.9 | 35.6 |
| | | MTL | 41.3 | 31.3 | **62.6** | **74.5** | 51.9 | 52.9 | 24.7 | 22.1 |
| 0.2 | BOW | STL | 54.0 | 60.3 | 59.3 | 68.9 | 56.6 | 64.6 | 30.1 | 42.0 |
| | | MTL | 53.6 | 51.1 | **64.0** | **74.2** | 58.8 | 62.6 | 32.4 | **38.5** |
| | BERT | STL | 48.2 | 54.5 | 57.5 | 69.0 | 52.8 | 61.7 | 25.4 | 35.0 |
| | | MTL | 46.5 | 39.4 | 62.4 | 72.2 | 54.5 | 55.8 | 25.7 | 24.8 |
| 0.0 | BOW | STL | 50.7 | 56.7 | 57.9 | 68.1 | 54.3 | 62.4 | 28.4 | 38.5 |
| | | MTL | 49.9 | 47.7 | 63.2 | 73.7 | 56.6 | 60.7 | 29.3 | 35.9 |
| | BERT | STL | **50.4** | **57.2** | **61.2** | 67.6 | **55.8** | **62.4** | **28.7** | **36.4** |
| | | MTL | **50.9** | **43.7** | 60.1 | 72.8 | **55.5** | **58.2** | **27.9** | **29.1** |

Table 10: Macro F1 scores for classification using MLP and different dropout rates: 0.5, 0.2, 0.0 (no dropout). For each task and setting, highest F1 scores for each combination of text representation and learning paradigm are presented in bold, absolute highest scores are underlined.