Policy Preference Detection in Parliamentary Debate Motions

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

Federico Nanni
Data and Web Science Group
University of Mannheim
federico@
informatik.uni-mannheim.de

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

Simone Paolo Ponzetto
Data and Web Science Group
University of Mannheim
simone@
informatik.uni-mannheim.de

Abstract

Debate *motions* (proposals) tabled in the UK Parliament contain information about the stated policy preferences of the Members of Parliament who propose them, and are key to the analysis of all subsequent speeches given in response to them. We attempt to automatically label debate motions with codes from a pre-existing coding scheme developed by political scientists for the annotation and analysis of political parties' manifestos. We develop annotation guidelines for the task of applying these codes to debate motions at two levels of granularity and produce a dataset of manually labelled examples. We evaluate the annotation process and the reliability and utility of the labelling scheme, finding that inter-annotator agreement is comparable with that of other studies conducted on manifesto data. Moreover, we test a variety of ways of automatically labelling motions with the codes, ranging from similarity matching to neural classification methods, and evaluate them against the gold standard labels. From these experiments, we note that established supervised baselines are not always able to improve over simple lexical heuristics. At the same time, we detect a clear and evident benefit when employing BERT, a state-of-the-art deep language representation model, even in classification scenarios with over 30 different labels and limited amounts of training data.

1 Introduction

Commonly known as the *Hansard* record, transcripts of debates that take place in the House of Commons of the United Kingdom (UK) Parliament are of interest to scholars of political science as well as the media and members of the public who wish to monitor the actions of their elected representatives. Debate *motions* (the proposals tabled for debate) are expressions of the policy positions taken by the governments, political parties, and individual Members of Parliament (MPs) who propose them. As all speeches given and all votes cast in the House are responses to one of these proposals, the motions are key to any understanding and analysis of the opinions and positions expressed in the subsequent speeches given in parliamentary debates.

By definition, debate motions convey the stated policy preferences of the MPs or parties who propose them. They therefore express polarity—*positive* or *negative*—towards some target, such as a piece of legislation, policy, or state of affairs. As noted by Thomas et al. (2006), the polarity of a debate proposal can strongly affect the language used by debate participants to either support or oppose it, effectively acting as a polarity shifter on the ensuing speeches. Analysis of debate motions is therefore a key first step in automatically determining the positions presented and opinions expressed by all speakers in the wider debates.

Additionally, there are further challenges associated with this task that differentiate it from the forms of sentiment analysis typically performed in other domains. Under Parliament's *Rules of Behaviour*,[1] debate participants use an esoteric speaking style that is not only laden with opaque procedural language and parliamentary jargon, but is also indirect, containing few explicitly negative words or phrases, even where negative positions are being expressed (Abercrombie and Batista-Navarro, 2018a).

The topics discussed in these debates revolve

[1] https://www.parliament.uk/documents/rules-of-behaviour.pdf

around policies and policy domains. Topic modelling or detection methods, which tend to produce coarse overviews and output neutral topics such as 'education' or 'transport' (as in Menini et al. (2017), for instance), are therefore not suitable for our purposes. Rather, we seek to find the proposer of a motion's position or *policy preference* towards each topic—in other words, an *opinion-topic*. Topic labels do exist for the Hansard transcripts, such as those produced by the House of Commons Library or parliamentary monitoring organsitions such as Public Whip.[2] However, these are unsuitable due to, in the former case, the fact that they incorporate no opinion or policy preference information, and for the latter, being unsystematic, insufficient in both quantity and coverage of the topics that appear in Hansard, and not future-proof (that is, they do not cover unseen topics that may arise (Abercrombie and Batista-Navarro, 2018b)).

In this paper, we use the coding scheme devised by the Manifesto Project,[3] because: *(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.

We approach automatic policy preference labelling at both the motion and (quasi-)sentence levels (see Section 2). We envisage that the output could therefore be used for downstream tasks, such as sentiment and stance analysis and agreement assessment of debate speeches, which may be performed at different levels of granularity.

Our contributions This paper makes the following contributions to the literature surrounding natural language processing of political documents and civic technology applications:

1. We develop a corpus of English language debate motions from the UK Parliament, annotated with policy position labels at two levels of granularity. We also produce annotation guidelines for this task, analysis of inter-annotator agreement rates, and further evaluation of the difficulty of the task on data from both parliamentary debates and the manifestos. We make these resources publicly available for the research community.

2. We test and evaluate two different ways of automatically labelling debate motions with Manifesto Project codes: lexical similarity matching and supervised classification. For the former, we compare a baseline of unigram overlap with cosine similarity measurement of vector representations of the texts. For the latter, we test a range of established baselines and state-of-the-art deep learning methods.

2 Background

Rather than being forums in which speakers attempt to persuade one another of their points of view, as the word 'debate' may imply, parliamentary speeches are displays of position-taking that MPs use to communicate their policy preferences to 'other members within their own party, to members of other parties, and, most important, to their voters' (Proksch and Slapin, 2015). Debate *motions* are proposals put forward in Parliament, and as such are the objects of all votes and decisions made by MPs, and, in theory at least, of all speeches and utterances made in the House.[4] Each parliamentary debate begins with such a motion, and may include further *amendment* motions (usually designed to alter or reverse the meaning of the original) as it progresses. Motions routinely begin with the words '*I beg to move That this House ...*', and may include multiple parts, as in *Example 1*,[5] which consists of two clauses, and appears to take a positive position towards international peace:

> *I beg to move*
> *That this House notes the worsening humanitarian crisis in Yemen;*
> *and calls upon the Government to take a lead in passing a resolution at the UN Security Council that would give effect to an immediate ceasefire in Yemen.* (1)

The concept of *policy preferences* is widely used in the political science literature (e.g. Budge

[2] https://www.publicwhip.org.uk
[3] https://manifestoproject.wzb.eu
[4] https://www.parliament.uk/site-information/glossary/motion
[5] https://hansard.parliament.uk/commons/2017-03-28/debates/F81005F8-5593-49F8-82F7-7A62CB62394A/Yemen

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 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 one of which (*408: Economic goals*) 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 included manifestos 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), as in *Example 2*, in which a single sentence has been annotated as four quasi-sentences:[6]

> *To secure your first job we will create 3 million new apprenticeships;*
> 411: Technology and Infrastructure
>
> *take everyone earning less than 12,500 out of Income Tax altogether*
> 404: Economic Planning
>
> *and pass a law to ensure we have a Tax-Free Minimum Wage in this country;*
> 412: Controlled Economy
>
> *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

(2)

3 Related work

There exists a large body of work concerning the analysis of opinions and policy positions in the related domains of legislative debate transcripts (for a survey, see Abercrombie and Batista-Navarro, 2019) and party political manifestos (see Volkens et al., 2015). Inspired by work on analysis of text from other domains, such as product reviews and social media, much of the computer science research in this area has concentrated on classifying the sentiment polarity of individual speeches (e.g. Burford et al., 2015; Thomas et al., 2006; Yogatama et al., 2015). Political scientists meanwhile, have tended to focus on position scaling—the task of placing the combined contributions of a political actor on a (usually) one-dimensional scale, such as *Left–Right* (e.g. Glavaš et al., 2017b; Laver et al., 2003; Nanni et al., 2019a; Proksch and Slapin, 2010). In either case, the majority of this work does not take into consideration the topics or policy areas addressed in the speeches.

Supervised classification approaches to opinion-topic identification have been explored in a number of papers. Abercrombie and Batista-Navarro (2018b) obtain good performance in classifying debate motions as belonging to one of 13 'policies' or opinion-topics. However, this approach is somewhat limited in that they use a set of pre-existing labelled examples which does not extend to cover the whole Hansard corpus or any new policies that may arise in the future. A similar setting to ours is that of Herzog et al. (2018), who use labels from the Comparative Agendas Project (CAP).[7] However, while they seek to discover latent topics present in the corpus, we wish to determine the policy-topic of each individual debate/motion. Rather than employ labelled manifesto data, as we do, they use the descriptions of the CAP codes.

Concerning policy identification in party political manifestos, previous studies have focused on topical segmentation (Glavaš et al., 2016) and classification of sentences into the seven coarse-grained policy domains (Glavaš et al., 2017a; Zirn et al., 2016). Meanwhile, Subramanian et al. (2018) recently presented a deep learning model that classifies manifesto sentences with the finer-grained code-level scheme of the Manifesto Project, as well as placing them on a Left-Right scale. In order to contribute to these research efforts and following recent advancements in deep language representation models (Devlin et al., 2018; Peters et al., 2018), we test the potential of BERT (Bidirectional Encoder Representations from Transformers) for policy-topic classification on both debate motions and manifestos.

[6] Conservative Party manifesto 2015.

[7] https://www.comparativeagendas.net

There is also a growing body of research on the evaluation of annotations for this domain. While the Manifesto Project relies on trained individual annotators to label manifestos, Mikhaylov et al. (2008) report the results of experiments which show that agreement between annotators is difficult to achieve, casting doubts on the reliability of the Project's codes. However, in similar experiments, Lacewell and Werner (2013) report greater inter-annotator agreement, and claim that with ongoing training, annotators can produce reliable labels. An extended analysis of the validity and reproducibility of the coding scheme is offered by Gemenis (2013), who remarks on the fact that 'the problem of unreliability does not lie with the coders but with the complex nature of the CMP (Comparative Manifesto Project) coding scheme'. Aware of such challenges, and in order to offer an additional comparison to these previous studies, in this work we provide a detailed analysis of the agreement rates of our annotators on both manifestos and debate motions.

4 Data

In the experimental section we report on the use of codes from the Manifesto Project as policy preference labels, with the goal of applying them to debate motions. These labels are convenient because: *(a)* like debate transcripts, they have been collected over time; and *(b)* the Project is ongoing, meaning that new example manifestos will continue to be added to it, mitigating potential concept drift problems (in which the language used to refer to aspects of different policy areas may change diachronically).

To construct our corpus, we made use of the data sources described below:

The Manifesto Project

We used annotated manifestos *(1)* as reference texts for labelling of debate motions by similarity matching, and *(2)* training a neural network for cross-domain classification of the motions. We downloaded all fifteen of the annotated United Kingdom (including Northern Ireland) manifestos 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.[8]

[8] https://manifestoproject.wzb.eu/coding_schemes/mp_v4

Domain	Manifestos	Debates	
	QSs	Motions	QSs
1: External Relations	1,436	50	186
2: Freedom & Democracy	767	30	106
3: Political System	1,627	47	220
4: Economy	4,296	87	380
5: Welfare & Quality of Life	2,235	118	528
6: Fabric of Society	1,574	33	153
7: Social Groups	1,180	21	110
0: No meaningful category	166	0	0

Table 1: The number of quasi-sentences (QSs) coded under each domain in the UK manifestos that we use as reference texts and training data and the number of debate motions and quasi-sentences that we label under each domain in the motion policy preference corpus.

Party	Year(s)	QSs
Conservative	2015	1589
DUP	2015	229
Green Party	2015	2235
Labour	2001, 2015	2503
Liberal Democrats	1997, 2015	2759
Plaid Cymru	2015	776
SDLP	2015	407
Sinn Féin	2015	272
SNP	1997, 2001, 2015	2309
UKIP	2015	1349
UUP	2015	417

Table 2: The parties and years of publication of the manifestos that we use as reference texts and training data, and the number of labelled quasi-sentences (QSs) by party in this subset of the manifesto data.

In this subset, the number of UK manifesto quasi-sentences labelled with codes in each domain varies considerably (see Table 1). These manifestos were written by a variety of political parties for elections over an 18 year period (Table 2). The most prevalent code in these manifestos is *504: Welfare State Expansion* (2,691 examples), and the least used is *103: Anti-Imperialism* (3 examples). Two codes, *102: Foreign Special Rela-*

tionships: Negative and *415: Marxist Analysis: Positive*, do not appear at all in manifestos from the United Kingdom.

Debate transcripts

The Hansard record of House of Commons debates is available for each day on which debates have taken place from 1919 to the present day in `xml` format at `https://www.theyworkforyou.com`, where it is updated daily with the most recent debates. As the record is more complete for recent years, we downloaded all files from May 7th 1997 (the start of that year's session of Parliament) to February 28th 2019. From these we extracted 1,156 motions together with the titles of the debates and the dates on which they were tabled. We 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 data in the Manifesto Project, and to investigate policy preference detection at different levels of granularity, we divided each motion into smaller units. For convenience, we approximated quasi-sentences in the Hansard data by automatically dividing motions into clauses, which are separated by semi-colons in the transcripts.

5 Annotation

We adapt the Project's *Coding Instructions* (Werner et al., 2011) to provide guidelines for the annotation of debate motions. We use 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, we 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. We used the coding instructions of version 4 of the Manifesto Project handbook[9] supplemented by debate motion-specific guidelines including notes based on the annotators' discussions during training.[10] 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 from 'fair' to 'substantial', following common interpretation of Fleiss' *kappa* scores (Landis and Koch, 1977) (see Table 3).

The final corpus includes 386 hand-annotated motions and 1,683 quasi-sentences.[11] 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 4). 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), the gold standard label for each example is obtained by adjudication, which was carried out by the first author.

Manifestos

To validate our labelling procedure, and for comparison with other work, we also asked the annotators to label a small quantity (120) of quasi-sentences from the Manifesto Project. We calculate 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 manifestos.

Again, we asked the annotators 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 ex-

[9] Available at `https://manifestoproject.wzb.eu/down/papers/handbook_2011_version_4.pdf`

[10] These guidelines are available along with the corpus.

[11] These constitute examples with 'gold standard' labels. The corpus also includes examples labelled by a sole annotator ('silver standard') and further unlabelled motions (see Table 5).

Iteration	Motion level			Quasi-sentence level		
	No. of examples	*k*	Interpretation	No. of examples	*k*	Interpretation
Training 1	15	0.41	'moderate'	60	0.35	'fair'
Training 2	12	0.65	'substantial'	60	0.56	'moderate'
Training 3	16	0.48	'moderate'	60	0.40	'fair'

Table 3: 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 codes from the Manifesto Project.

	Annotators	No.	*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
QS	1 & 2	622	0.58
	2 & 3	650	0.51
	1 & 3	731	0.62
	All pairs	2003	0.58

Table 4: Fleiss' *kappa* scores for three-way agreement and Cohen's *kappa* scores for two-way agreement on the debate motions dataset.

Labels	Motion	Quasi-sentence
Gold standard	386	1,683
Silver standard	87	361
Total labelled	473	2,044
Unlabelled	593	2,587
Overall total	1,066	4,631

Table 5: 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.

ample marked as such by all three annotators. In this case, two of them used the 'correct' Manifesto Project gold label, while the third annotator applied a different code from the same domain. Overall, 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.

The Motion Policy Preference Corpus

We make the corpus 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 5. 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).

6 Automatic Labelling Methods

We investigated two ways of automatically labelling debate motions with the codes from the Manifesto Project: *(1)* similarity matching and *(2)* supervised classification. We tested both at the quasi-sentence level and we additionally experiment 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 we 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.

Similarity matching

We 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, we 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, we simply counted the union of the sets of tokens from *A* and *B*. For the latter method, each text was represented by its term frequency-inverse document frequency vector (tf-idf), and cosine similarity calculated as:

$$\frac{\vec{A} \cdot \vec{B}}{||\vec{A}||||\vec{B}||}$$

With both of these approaches, we explored the use of the following combinations of sources of textual unigram features: the debate titles, which have been shown to be highly predictive of a

motion's opinion-topic in a supervised classification setting (Abercrombie and Batista-Navarro, 2018b), the debate motions themselves, and both the titles and motions together.

Supervised Classification

We 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. We were particularly interested in assessing the performance of such approaches: *(1)* despite the limited training data available (1.6k motion quasi-sentences); and *(2)* in a cross-domain application (training on over 16k manifesto quasi-sentences, and testing on the motion quasi-sentences).

First, we examined the performance of Support Vector Machines (SVM) trained using lexical (tf-idf) or word embedding (w-emb) features, which act as strong traditional baselines. We tested both pre-trained general purpose word embeddings from `https://fasttext.cc` (Mikolov et al., 2018) and in-domain vectors generated on the Hansard transcripts from Nanni et al. (2019b).

We 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 Convolutional Neural Network (CNN) with single convolution layer and a single max-pooling layer. We again tested the CNN with general purpose and in-domain embeddings.

As final skyline comparisons, we present the performance of *(1)* a pre-trained BERT (large, cased) model (Devlin et al., 2018), with a final soft-max layer; and *(2)* the same pre-trained BERT model, with a CNN and max-pooling layers before the soft-max layer. We additionally experimented with the latter two models in a fine-tuning setting: after training on manifestos, they have been further fine-tuned on motions.

We tested all approaches with a 80/20 split of the dataset, and trained all the neural models for three iterations.

7 Results

We evaluated the predicted labels of each experimental model against the gold standard labels produced by the annotation process. For the machine learning methods, we 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

We 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 the policy and domain levels (see Figure 1).

Unlike the findings of Abercrombie and Batista-Navarro (2018b), in most settings, we do 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 our larger set of class labels containing more similar (same domain) policy preference codes.

Best performances at both policy and domain levels (F1 macro = 0.59) are obtained using 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.

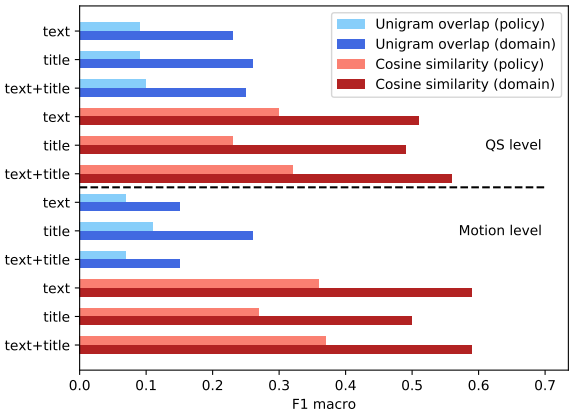

Figure 1: F1 macro scores for unigram overlap and cosine similarity matching at the policy and domain levels using textual features from whole motions. Use of cosine similarity leads to markedly better performance than unigram overlap, and the best performance is achieved using features derived from both the titles and motion texts at policy and domain levels.

Motions: Quasi-sentence Classification

We tested the supervised pipelines at the quasi-sentence level and at the two levels of class label granularity (policy and domain), which allows

Model	Text representation	Data Source	Policy		Domain	
			Macro	Micro	Macro	Micro
Unigram overlap	BOW	Motion titles	0.09	0.15	0.26	0.31
		Motions	0.09	0.21	0.23	0.38
		Titles+motions	0.10	0.23	0.25	0.39
Cosine similarity	Tf-idf	Motion titles	0.23	0.34	0.44	0.49
		Motions	0.30	0.36	0.50	0.51
		Titles+motions	0.32	0.41	0.51	0.56
SVM	Tf-idf	Motions	0.33	0.48	0.58	0.63
		Manifestos	0.29	0.40	0.53	0.56
	Domain w-emb	Motions	0.32	0.50	0.53	0.62
		Manifestos	0.25	0.41	0.45	0.53
	Wiki w-emb	Motions	0.35	0.51	0.55	0.65
		Manifestos	0.21	0.38	0.45	0.52
CNN	Domain w-emb	Motions	0.15	0.38	0.58	0.64
		Manifestos	0.19	0.30	0.37	0.51
	Wiki w-emb	Motions	0.13	0.29	0.50	0.57
		Manifestos	0.21	0.36	0.48	0.56
BERT	Large, cased	Motions	0.26	0.47	0.42	0.58
		Manifestos	0.32	0.47	0.52	0.57
		+ Motions fine-tuning	0.39	0.50	0.60	0.67
BERT+CNN	Large, cased	Motions	0.27	0.48	0.42	0.56
		Manifestos	0.29	0.44	0.54	0.60
		+ Motions fine-tuning	**0.47**	**0.57**	**0.61**	**0.69**

Table 6: F1 scores for similarity matching and classification of debate motions at the quasi-sentence level.

us to compare the results with previous work on the Manifesto Project (e.g., Zirn et al. (2016)). As can be seen in Table 6, 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, we notice 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 the employed resources 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, we return to this issue.

Finally, to further confirm the large potential of BERT, even in tasks which involve many labels, a lack of training data, and a very specific style of communication, we have obtained a clear improvement over all other systems when employing this state-of-the-art architecture, trained on manifesto quasi-sentences and further fine-tuned on motions.

Manifestos: Quasi-sentence Classification

As a final comparison of the presented systems for quasi-sentence classification, we report their performance on the corpus of 16k manifesto quasi-sentences, again with an 80/20 train-test split. The results (see Table 7) 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 we obtained on the motion corpus in Table 6.

Interestingly, we once again notice the weak performances of the CNNs 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. Con-

Model	Text representation	Policy		Domain	
		Macro	Micro	Macro	Micro
SVM	Tf-idf	0.39	0.54	0.58	0.66
	Domain w-emb	0.35	0.53	0.52	0.64
	Wiki w-emb	0.38	0.54	0.54	0.66
CNN	Domain w-emb	0.28	0.47	0.54	0.58
	Wiki w-emb	0.27	0.44	0.52	0.56
BERT	Large, cased	**0.42**	**0.58**	0.58	0.64
BERT + CNN	Large, cased	**0.42**	**0.58**	**0.60**	**0.70**

Table 7: F1 scores for classification of party political manifestos at the quasi-sentence level.

versely, traditional SVM baselines offer reasonable results, and we achieve state-of-the-art performances when employing BERT.

8 Discussion and Conclusion

Through this work we 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, our 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[12] itself (Lacewell and Werner, 2013).

The task does seem to be transferable to parliamentary debate motions, with our inter-annotator agreement scores comparable on both domains. Although automatic labelling with lexical similarity matching is more succesful at the quasi-sentence level than at the motion level, the annotators do not seem to find the coarser grained 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 find 230 articles that use this data in the eight journals they examine. With comparable reliabilty (inter-annotator agreement), the labelled motions could prove equally suitable for many automatic analysis applications.

Concerning automation of the labeling process, we can 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 our supervised pipelines, which are able to achieve state-of-the-art performance on both the motions and manifesto quasi-sentences.

The generated dataset of topically labelled motions along with the trained BERT+CNN classifier can now pave the way for further work at the intersection of natural language processing and political science, which can benefit from these fine-grained policy position annotations: from analysing the sentiment of the motions to measuring the level of disagreement between members of the same party, and up to full-blown argumentation mining of each debate.

[12]Manifesto Research on Political Representation, the research team behind the Manifesto Project.

Acknowledgements

This work was supported in part by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (projects B6 and C4), funded by the German Research Foundation (DFG). The authors would like to thank Melis Ince, Olga Sokolova, and Stefan Tasic for their diligent work on annotation, and the anonymous reviewers for their helpful comments.

References

Gavin Abercrombie 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).

Gavin Abercrombie and Riza Batista-Navarro. 2019. Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *arXiv preprint arXiv:1907.04126*.

Gavin Abercrombie 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.

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

Clint Burford, Steven Bird, and Timothy Baldwin. 2015. Collective document classification with implicit inter-document semantic relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 106–116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kostas Gemenis. 2013. What to do (and not to do) with the Comparative Manifestos Project data. *Political Studies*, 61:3–23.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017a. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science (NLP+CSS)*, pages 42–46.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017b. Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693.

Alexander Herzog, Peter John, and Slava Jankin Mikhaylov. 2018. Transfer topic labeling with domain-specific knowledge base: An analysis of UK House of Commons speeches 1935-2014. *arXiv preprint arXiv:1806.00793*.

Werner Krause, Pola Lehmann, Jirka Lewandowski, Theres Matthie, Nicolas Merz, Sven Regel, and Annika Werner. 2018. Manifesto Corpus. version: 2018-1. Berlin: WZB Berlin Social Science Center.

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

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

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling policy preferences from coded political texts. *Legislative Studies Quarterly*, 36(1):123–155.

Stefano Menini, 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.

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

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Federico Nanni, Goran Glavaš, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2019a. Political text scaling meets computational semantics. *arXiv preprint arXiv:1904.06217*.

Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019b. Semantifying the UK Hansard (1918-2018). In *Proceedings of the Joint Conference on Digital Libraries*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Sven-Oliver Proksch and Jonathan B Slapin. 2010. Position taking in European Parliament speeches. *British Journal of Political Science*, 40(3):587–611.

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

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1964–1974.

Matt Thomas, 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. Association for Computational Linguistics.

Andrea Volkens, 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.

Andrea Volkens, Judith Bara, Ian Budge, Michael D McDonald, and Hans-Dieter Klingemann. 2013. *Mapping policy preferences from texts: statistical solutions for manifesto analysts*, volume 3. OUP Oxford.

Annika Werner, Onawa Lacewell, and Andrea Volkens. 2011. Manifesto coding instructions: 4th fully revised edition.

Dani Yogatama, Lingpeng Kong, and Noah A. Smith. 2015. Bayesian optimization of text representations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2100–2105, Lisbon, Portugal. Association for Computational Linguistics.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos. In *Proceedings of the First International Conference on the Advances in Computational Analysis of Political Text (PolText)*.