PTPARL-V: Portuguese Parliamentary Debates for Voting Behaviour Study

Afonso Sousa, Henrique Lopes Cardoso
Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto, Portugal
{ammlss,hlc}@fe.up.pt

Abstract

We present a new dataset, PTPARL-V, that is a valuable resource for advancing discourse analysis of parliamentary debates in Portuguese and their alignment with voting behaviour. This is achieved by processing the open-access information available at the official Portuguese Parliament website and scraping the debate minutes concerning legislative initiatives, together with meta-data related to voting positions. Our dataset includes interventions from 547 different deputies of all major Portuguese parties, from 736 legislative initiatives spanning five legislatures from 2005 to 2021. We present a statistical analysis of the dataset compared to other publicly available Portuguese parliamentary debate corpora. Finally, we provide baseline performance analysis for voting behaviour classification.

Keywords: Portuguese debates, Discourse analysis, Parliamentary data, Voting behaviour

1. Introduction

Parliamentary corpora are essential language resources that can be approached from various research perspectives, including political science, sociology, history, and psychology. Parliamentary and legislative debate transcripts provide access to information concerning elected politicians' opinions, positions, and policy preferences. This kind of information can be used for a variety of computational tasks and natural language applications, such as critical discourse analysis (Van Dijk, 1993), sentiment analysis (Abercrombie and Batista-Navarro, 2020), argument detection (Cabrio and Villata, 2018) or stance detection (Schiller et al., 2021).

The digitization of parliamentary documents and the advancement of computer tools have created interesting opportunities for political data analysis. In recent years, efforts have been made to compile well-structured corpora of parliamentary debates. At the time of writing, the CLARIN infrastructure offers access to 35 parliamentary corpora[1], covering most languages spoken in European countries.

However, almost all corpora are solely comprised of text passages and tags extracted from postprocessing said text (e.g., POS tagging or NER). Moreover, large-scale research on voting discipline and behaviour does not compare discourses, instead solely focusing on the scattering of votes through the various parties (Kam, 2009).

There are many compilations of parliamentary debates. To our knowledge, specifically for Portuguese there is PTPARL (Généreux et al., 2012), a compendium of Portuguese parliamentary debates from 1970 to 2008; PTPARL-D (Almeida et al., 2021), a compilation of all debates of the Third Portuguese Republic, spanning 44 years; and another speech compilation (Fernandes et al., 2021) comprised of speeches from 1999 to 2017. None of these includes annotations for any NLP task.

We introduce PTPARL-V, a new Portuguese dataset that addresses the voting behaviour of members of the Portuguese Parliament. We compiled interventions across five legislatures and extracted associated metadata. We gathered information on the initiatives voted in favour, against, or abstained by all major parties in the Portuguese Parliament. We expect this work to help produce more thorough studies regarding voting behaviour from the different parties and their members.

In summary, the contributions of this paper are: (i) a new Portuguese dataset for voting behaviour analysis of political debates; (ii) a statistical analysis of the newly created dataset; and (iii) a preliminary baseline performance benchmark for forecasting voting behaviour.[2]

2. About Legislative Initiatives

The Portuguese Parliament (*Assembleia da República*) provides open-access data on parliamentary activities on its official website[3]. We next describe the source of information and how the interventions are selected.

Many different activities are conducted in parliament. We focus on *legislative initiatives*: proposals for new laws. These initiatives can be proposed

[1] https://www.clarin.eu/resource-families/parliamentary-corpora

[2] The dataset and code were made available at https://github.com/afonso-sousa/pt_parliamentary_minutes

[3] https://www.parlamento.pt/

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 38–42

by members of the parliament (MPs), parliamentary groups or groups of voting citizens – draft law (*projeto de lei*) – or by the Government or the Regional Legislative Assemblies (RLA) – proposed law (*proposta de lei*). After being admitted by the President of the Assembly, the initiative is subjected to an assessment by the specialised Commission to which it has been assigned, followed by its general debate in a plenary meeting, which ends with a voting process. Further steps may be taken for an initiative to be considered law. For voting behaviour and discourse analysis, we built our dataset by collecting the plenary debate and the general voting information. We discarded joint initiatives (multiple initiatives discussed in the same plenary meeting) because the respective transcripts are cluttered with different subjects and themes, making their automatic parsing and clear distinction of initiatives unfeasible. These initiatives are published in *Diário da República* – the official Portuguese journal where laws, decisions by the Constitutional Court and other relevant texts are published.

The represented parties in *Assembleia da República* that intervened in the plenary meetings to discuss the initiatives mentioned above are briefly summarised in Table 1.

3. Dataset Compilation

We next describe what attributes were selected and how the PTPARL-V dataset was built.

While the open-access data is available in common formats, like XML or JSON, processing the free-text concerning MP speeches is not trivial, as these are contained within PDF files embedded in the website:

- We first downloaded all the published transcripts matching our time span: legislatures X to XIV, spanning from 2005 to 2021.
- Then, from the open-access data, we collected initiatives that matched our previously settled requirements: legislative initiatives (avoiding joint ones) with plenary debate and a general vote. From these, we collected relevant attributes to characterize an entry in the dataset.
- We extracted the text from the transcripts related to the collected initiatives. Each initiative has annotations of the pages within *Diário da República* with the discussion of the initiative. We used text extraction tools to retrieve the text from the designated PDF pages.
- From the retrieved pages, we matched the deputy's name in the metadata with the speaker's name at the beginning of the paragraph (see bold in Figure 1) and concatenated the collected paragraphs. This step produces a multi-sentence text passage comprised of all paragraphs of a deputy's speeches in the discussion of the initiative.

The above-mentioned steps produce text about each MP's stance towards a given initiative – an *intervention*. This information, along with the corresponding metadata, makes up an entry in the dataset. The metadata serves to characterise the intervention and covers three main concepts: the intervention, the initiative for which the intervention was made, and the legislature in which the initiative was proposed. The *intervention* is made by an MP, who has a name and a party they belong to. The intervention also has information on the MP's vote on the initiative being discussed: in favour, against or abstention. The *initiative* has information about the proponents, the type of initiative, and a summary description of the topics being discussed. Lastly, the initiative is proposed in a given legislative session within a *legislature*, identified by a Roman numeral and temporally framed.

4. Data Analysis

We analyse some properties of our dataset.

4.1. Basic Statistics

In Table 2, we compare basic statistics between PTPARL (Généreux et al., 2012) and PTPARL-V. After cleaning, PTPARL-V has a total of 736 initiatives and 5833 interventions (see Table 3 for a distribution over legislatures). To the best of our knowledge, PTPARL is the only previously publicly available compilation of interventions in the Portuguese parliament. PTPARL-V is much larger than PTPARL, with the added benefit of having the accompanying metadata (including voting behaviour).

As for general metadata statistics, Table 4 shows some overall information on per-party initiatives and interventions. There are approximately 10 interventions per party per initiative.

4.2. Exploratory Data Analysis

From the metadata alone, we can judge the political scene in Portugal for the dataset time frame. By aggregating similar votes for each initiative, Figure 2 shows the eight sets of parties with the highest similar vote frequency. This means that if an initiative was voted in favour by, say, both PSD and CDS-PP, it would count +1 towards the 'in favour' bar of the "PDS,CDS-PP' set. From the plot, we see that parties often vote in favour of the proposed initiatives, as given by the overall higher frequencies in the respective bars. Additionally, we can see that parties closer in the political spectrum (namely PSD and CDS-PP, or BE, PCP and PEV, see Table 1)

Party Initials	Full Name	Main Ideology	Position
PCP	Portuguese Communist Party, *Partido Comunista Português*	Marxism-Leninism	Left-wing to far-left
BE	Left Bloc, *Bloco de Esquerda*	Democratic socialism	Left-wing to far-left
PEV	Ecologist Party "The Greens", *Partido Ecologista "Os Verdes"*	Eco-socialism	Left-wing
PS	Socialist Party, *Partido Socialista*	Social Democracy	Centre-left
PAN	People Animals Nature, *Pessoas-Animais-Natureza*	Environmentalism	Centre-left
PSD	Social Democratic Party, *Partido Social Democrata*	Liberal conservatism	Centre-right
CDS-PP	Democratic and Social Centre - People's Party, *Centro Democrático e Social – Partido Popular*	Conservatism	Centre-right to right-wing
IL	Liberal Initiative, *Iniciativa Liberal*	Classical liberalism	Centre-right to right-wing
CH	ENOUGH, *CHEGA*	Right-wing populism	Right-wing to far-right

Table 1: General information on the Portuguese parties that have/had representation in *Assembleia da República* (retrieved from Wikipedia).

A Sr.ª **Mariana Aiveca** (BE): — Combater a precariedade e os falsos «recibos verdes», acabar com práticas de contratação ilegal criminalizando os seus responsáveis é o objectivo principal do projecto de lei que trazemos hoje a debate.

Figure 1: Sample paragraph from an intervention in *Diário da República*.

Dataset	# tokens	# sentences
PTPARL	975 806	48 911
PTPARL-V	3 790 086	111 614

Table 2: Basic dataset statistics for PTPARL-V and PTPARL (retrieved from PORTULAN Clarin).

legislature	# initiatives	# interventions
X	211	1609
XI	46	416
XII	267	2061
XIII	152	1239
XIV	60	508
Total	736	5833

Table 3: Distribution of initiatives and interventions per legislature in the PTPARL-V dataset.

Party	# initiatives	# votes (favour/against/abst)
Government	443	–
RLA Madeira	29	–
RLA Açores	13	–
PCP	38	512/357/162
BE	59	529/297/146
PEV	21	135/90/36
PS	38	726/367/101
PAN	4	45/7/13
PSD	32	642/336/247
CDS-PP	31	473/293/237
IL	0	12/9/4
CH	0	17/12/8
Mixed	27	–
Citizens	1	–

Table 4: Initiatives and intervention votes in the PTPARL-V dataset. 'Mixed' refers to initiatives authored by deputies of different parties.

often vote together. While a centre-left party, PS is often seen voting alone, explained by the fact that PS is often the governing party, sometimes with an absolute majority.

In Figure 3, we see a distribution of initiatives and interventions per year. Legislatures X and XII were the ones where more initiatives were proposed, spanning from 2006 to 2009 and from 2011 to 2015, respectively.

5. Predicting Voting Behaviour

In this section, we model the task of predicting voting behaviour as a supervised multiclass classification problem. We try to predict if a given speaker will be voting 'in favour', 'against', or 'abstaining' based on the contents of their interventions.

We randomly split the dataset into train and test

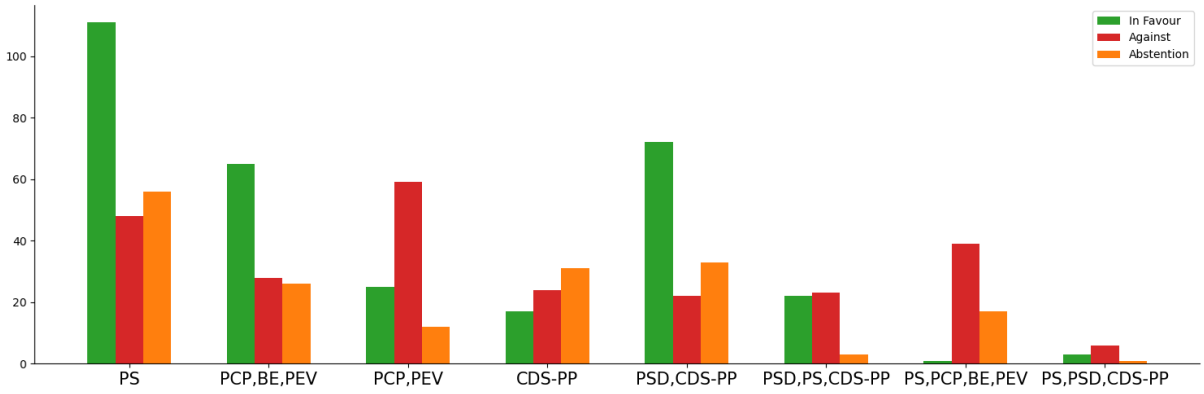

Figure 2: *Who votes with whom?* These bar charts show the parliamentary sets of parties that most frequently voted together. This data was compiled using the frequency of initiative votes for every combination of parties in the dataset.

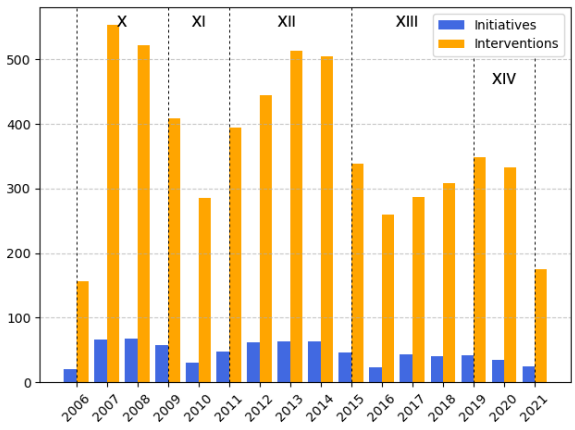

Figure 3: Initiatives and interventions per year.

sets in a stratified fashion (i.e., we keep the original distribution of labels in each set). The splits contain roughly 80% and 20% of the total entries, respectively. We trained a Naive Bayes classifier with TF-IDF features and a Logistic Regression classifier with word embedding features. We also fine-tuned a pretrained multilingual DistilBERT (Sanh et al., 2019)[4] model, the Portuguese encoder BERTimbau (Souza et al., 2020)[5] base model, and two versions of ALBERTINA (900M and 1.5B parameters versions[6], the latter being fine-tuned with LoRA (Hu et al., 2022)).

For feature-based models (Naive Bayes and Logistic Regression), we preprocessed the data:

- We removed all special characters.
- We converted all the text to lowercase.
- We removed generic stopwords (e.g., determiners, conjunctions and prepositions) and domain-specific stopwords (e.g., 'Sr.' (Mr.), 'secretário' (secretary), etc.). These domain-specific words are very prevalent in this compilation of parliamentary debates since there exists a standardized introductory etiquette for addressing the assembly.

The word embeddings used for the Logistic Regression were FastText CBOW with 50 dimensions[7] and were averaged to produce document-level embeddings. We relied on scikit-learn's[8] implementations of the feature-based models.

Model	Accuracy	Precision	Recall	F1
Naive Bayes	0.5904	0.4004	0.4267	0.3934
Logistic Regression	0.5687	0.5012	0.4189	0.3908
DistilBERT	0.5915	0.3830	0.4480	0.4123
BERTimbau	**0.6075**	**0.5240**	**0.4739**	**0.4711**
ALBERTINA-900M	0.5409	0.4416	0.3962	0.3790
ALBERTINA-1.5B-LoRA	0.4989	0.2363	0.3333	0.2639

Table 5: Multiclass classification performance on PTPARL-V.

Looking at Table 5, we find that all models can produce results better than a majority baseline for our dataset, that is, given the three-class split of our dataset is around 53-30-17, respectively for in-favour, against and abstention labels (see Figure 4 for details), the performance of our models is superior to just predicting the majority class, which would give an accuracy of around 53%. As such, we can assume some knowledge is contained within the text passages that can convey the voting behaviour of the speakers.

Interestingly, larger models did not perform better than the 110M parameter BERTimbau. We address this to the somewhat limited amount of training samples. Other issues with the dataset and models

[4] https://huggingface.co/distilbert/distilbert-base-multilingual-cased

[5] https://huggingface.co/neuralmind/bert-base-portuguese-cased

[6] https://huggingface.co/PORTULAN

[7] http://nilc.icmc.usp.br/embeddings

[8] https://scikit-learn.org/stable/supervised_learning.html

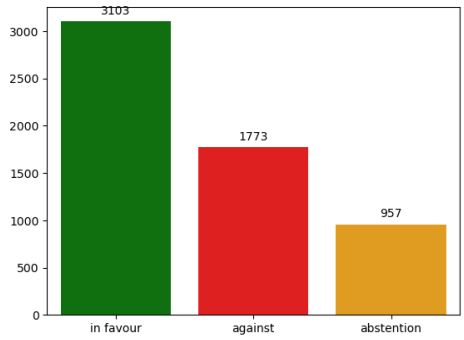

Figure 4: Number of instances per category.

used are as follows. The data contains repetitive passages of formally addressing the President and utterances of disagreement that do not contribute to the argument being made. Additionally, the texts are too long for the 512-token context cap of all the transformer-based models we tested. For reference, this dataset has an average word count of 712 words, meaning nearly half of the text is truncated. We believe a more careful consideration or text preprocessing will alleviate this issue.

6. Discussion and Conclusion

We introduced a new dataset, PTPARL-V, built from interventions of MPs in the Portuguese Parliament for six legislatures. We also briefly show the potential of such a dataset for political debate analysis – with some examples from exploratory data analysis showing the behavioural patterns of voting in the Portuguese Parliament – and vote behaviour forecasting – with a baseline classifier for vote prediction. Future improvements may still be made to the dataset. As for political debate analysis, we just scratched the surface of the insights that can be uncovered from a dataset like this, so we encourage anyone using this dataset to further the research on the Portuguese political scene. Finally, for vote prediction, thoroughly cleaning the text passages can significantly improve the performance of the classifiers. Additionally, using argument mining may be an interesting direction to uncover the most relevant discourse units that best indicate the voting preferences. The dataset and code to reproduce results were made available.

Acknowledgments

Afonso Sousa is supported by a PhD studentship (reference 2022.13409.BD), funded by Fundação para a Ciência e a Tecnologia (FCT). This research was supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC).

Bibliographical References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Paulo Almeida, Manuel Marques-Pita, and Joana Gonçalves-Sá. 2021. Ptparl-d: an annotated corpus of forty-four years of portuguese parliamentary debates. *Corpora*, 16(3):337–348.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Jorge M Fernandes, Mariana Lopes da Fonseca, and Miguel Won. 2021. Closing the gender gap in legislative debates: The role of gender quotas. *Political Behavior*, pages 1–25.

Michel Généreux, Iris Hendrickx, and Amália Mendes. 2012. A large portuguese corpus online: Cleaning and preprocessing. In *Computational Processing of the Portuguese Language*, pages 113–120, Berlin, Heidelberg. Springer Berlin Heidelberg.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Christopher J Kam. 2009. *Party discipline and parliamentary politics*. Cambridge University Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, 35(3):329–341.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Teun A Van Dijk. 1993. Principles of critical discourse analysis. *Discourse & society*, 4(2):249–283.