Information Extraction on Policy Preferences from Party Manifestos, Political Speeches and Opinion Articles

Nuno Amaro[1 2]
[1] INESC-ID
[2] Instituto Superior Técnico, University of Lisbon
nuno.amaro@ist.utl.pt

Abstract

This article presents an approach based on multilingual neural models for analysing political statements within textual documents, specifically considering a pipeline of two tasks: (a) identifying spans of text presenting political statements (i.e., relevant phrases describing policy positions), and (b) classifying the spans containing political statements according to a fine-grained set of classes derived from the codebooks of the Comparative Agendas and Manifesto projects. Both these tasks correspond to challenging problems that have seldom been addressed within the NLP or IR communities, e.g. due to semantic ambiguity or scarcity of proper training resources. The proposed approach was evaluated through experiments on Portuguese, Brazilian, Italian, and Spanish texts, collected from the aforementioned projects. Qualitative results are also shown regarding the analysis of opinion articles collected from Portuguese online newspapers. The obtained results suggest that automated classification is indeed viable for large scale computational social sciences studies.

Introduction

The manual coding of political texts is a difficult, long, and monotonous process, often undertaken as a pre-processing step for other downstream endeavours within social sciences studies (Merz, Regel, and Lewandowski 2016). In many instances, it may be necessary for researchers to read over several pages of political rhetoric and carefully annotate where a topic has shifted. Considering that some political texts, such as manifestos, can easily exceed hundreds of pages, large-scale endeavours to manually annotate topics within these contents are extremely time consuming and, therefore, this analysis is a natural candidate for automation.

In our work, leveraging developments in multilingual text modeling (Ruder, Vulić, and Søgaard 2019; Liu, Kusner, and Blunsom 2020), we present a multilingual approach for the recognition and classification of relevant text spans found within political texts. We specifically fine-tune state-of-the-art pre-trained contextual language models, based on the Transformer neural architecture, on both of these objectives (i.e., recognitiona and classification) and using data from the Comparative Agendas and Manifesto projects. Leveraging the individual models from each task, we also present a pipelined approach that is able to receive text, separate it into context spans, if needed, and classify the spans according to a fine-grained taxonomy. Our work also explored techniques for improving model performance, such as in-domain language model fine-tuning prior to the use in the recognition and classification tasks, or the augmentation of the available training data through machine translation. We argue the the models developed in this work can support the analysis of different types of political text and, to support this claim, we present qualitative results regarding the analysis of opinion articles collected from Portuguese online newspapers. The obtained results suggest that automated content analysis is a viable approach for large scale computational social science studies. The datasets and source code supporting the experiments will be released in a public GitHub repository.

Related Work

Our work uses data made available through two public sources of labeled political texts, namely the Manifesto Project[1] (MAN) and the Comparative Agendas Project[2] (CAP). In particular, MAN is a long running initiative that aims to subdivide (identify) political manifestos into different spans of text that are then classified using their coding scheme and instructions. In turn, CAP has a broader scope, covering several different forms of policy reporting activities (e.g. newspaper headlines, hearings, party manifestos, social media posts, etc.) under a, generally, consistent coding system. Both projects rely on domain experts for manually coding documents, which in many cases is a very time consuming task. These projects have allowed the academic community to study the evolution of policies and the importance of certain themes over long periods of time.

In the NLP and IR communities, some previous studies have addressed tasks related to classifying political texts, in some cases using data from the aforementioned projects. For instance, Karan et al. (2016) described the classification of titles taken from a series of different documents from the Croatian government and parliament, including the original manual coding of texts using the CAP scheme. These authors used classical feature-based text classification methods, such as logistic regression or naïve Bayes. Another pre-

[1] https://manifestoproject.wzb.eu/
[2] https://www.comparativeagendas.net/

	Top-1	Top-2	Top-3	Top-4	Top-5
Correct matches found	143	162 (+19)	174 (+12)	183 (+9)	188 (+5)
Percentage of correct matches	67.14	76.06	81.69	85.92	88.26

Table 1: Results for manual assessment of the categories retrieved for each CAP minor topic, out of a total of 213 CAP minor topics that were mapped into 46 MAN categories.

vious study (Subramanian et al. 2017) used shallow neural networks to classify text spans at a sentence-level, afterwards using these results in a hierarchical neural network for a regression task at the document level. While the document-level regression task focuses in attributing a political spectrum (right-left) leaning index to an entire manifesto, the sentence-level classification is much closer to our objective, as the authors proposed to classify each sentence into one of the MAN categories. Similarly to our work, Subramanian et al. (2017) also proposed to process text in multiple languages using multilingual word embeddings (Ammar et al. 2016). In subsequent work (Subramanian, Cohn, and Baldwin 2018), this classification approach would be further refined using neural models based on bidirectional LSTMs.

None of the aforementioned studies has considered the identification of the relevant text spans, instead focusing solely on classifying them. Our work is also novel in the use of state-of-the-art contextual language models.

Annotated Datasets of Political Rhetoric

This section introduces the datasets used in our study. We specifically used data from both the CAP and MAN, mapping the codebooks from these projects into a single taxonomy inspired by a reduced version of the MAN category system. Our data sample considers political manifestos for both presidential and legislative (where available) elections from three countries (i.e., Portugal, Brazil and Italy), and newspaper headlines from Spain.

In total, we use text from four different countries and spanning three similar languages, although we only use the newspaper headlines for the classification of text spans (i.e., data from long multi-sentence documents are used both for span identification and classification, whereas the news headlines are assigned entirely to a single class, and hence they are used only for span classification).

Leveraging the CAP and MAN data, we performed quantitative tests to assess the performance of the automatic models for relevant span identification and classification, and we also qualitatively analyzed the results of applying these models to the analysis of other types of text (e.g., opinion articles collected from Portuguese online newspapers).

Taxonomy Matching Using Document Similarity

Independent sources generally encode their data using different taxonomies, thus giving rise to the need for reducing encodings to a single taxonomy when attempting to combine/integrate contents. While this could be achievable using some domain knowledge of each category in order to manually find correspondences between taxonomies, we decided to adopt a faster and more automated approach for mapping the codebooks from the CAP and MAN projects. We specifically explored the possibility of matching categories using their respective textual descriptions, available in both codebooks, corresponding to a document similarity task. For the purposes of our work, we decided to map the over 200 CAP minor topics into one of the MAN categories, which is around a quarter of the size. We also reduced the total MAN scheme to a total of 46 three digit categories, by using version 4 of their codebook (which does not contain subcategories) and collapsing similar categories involving a different sentiment into a single category. All MAN categories belong to one of eight high-level domains corresponding to a coarser level of the codebook, which can be easily obtained by taking the left-most digit of the category.

An initial pre-processing of the textual descriptions was conducted, including removal of stop words and non-important parts through simple regular expressions (e.g., sentiment-specific language). We then utilized a publicly available pre-trained *word2vec* (Mikolov et al. 2013) model, which was originally trained on a large Google News dataset (about 100 billion words), containing 300-dimensional vectors for 3 million words and phrases. These vectors are L2 normalized before being used to compute the Word Mover's Distance (Kusner et al. 2015) between descriptions of each CAP minor topic towards all MAN categories. The five most similar categories for each CAP minor topic were then included in a spreadsheet for manual verification. This last step was conducted by first assigning one of three different labels (Good, Neutral and Bad), followed by a second verification. While certain topics can have more than one possible correspondence, we only consider the first (best) category found in the top 5. Categories where a *Good* match was not found through this process were paired manually through the analysis of the 46 MAN categories. Table 1 presents how many correct matches were found within of the top five categories retrieved through the Word Mover's Distance.

Datasets for Span Identification and Classification

Table 2 contains summary information on the part of the dataset that can be used to support experiments related to the identification of relevant textual spans, not including the validation split. This dataset is composed of full manifestos, separated by sentences, which can have zero, one or more relevant spans, each with its own class. The dataset contains all sentences from the materials used for span classification, which are in turn presented in Table 3, except for the Spanish contents which consisted of individual news headlines. The complete manifestos were split into training, testing and validation subsets, approximating the distribution of 70%, 15%

	Train			Test		
Language	Sentences	Relevant Spans	Avg. Tokens	Sentences	Relevant Spans	Avg. Tokens
Portuguese	11323	14470	33	7363	7852	29
Brazilian Portuguese	29468	35725	28	1974	2643	28
Italian	10159	11178	30	1315	1657	35
Total	50950	61373	29	10652	12152	30

Table 2: Description of dataset used in the tests related to span identification. Note that we have not included any Spanish data, as the available texts correspond to individual news headlines and not multi-sentence texts, which would be required for span identification. The columns named *Avg. Tokens* report the average number of tokens per sentence.

		Number of Spans				
Span Type	MAN Codes	PT	BR	IT	ES	Total
Relevant	101, 103, 104, 106, 107, 108	1589	2187	1337	12102	17215
	201, 202, 203	921	2913	1018	3819	8671
	301, 302, 303, 304, 305	3485	4405	3035	19557	30482
	401-416	7699	12396	5081	17891	43024
	501, 502, 503, 504, 506	8721	11682	5788	29182	55373
	601, 603, 605, 606, 607	1714	2640	1877	21247	27478
	701, 703, 704, 705, 706	1977	3946	1640	1967	9530
Irrelevant	000, 999	2813	3493	2407	1407	10120
Total		28919	43662	22183	107172	201936

Table 3: High-level description of the main dataset used for span classification. This dataset is based of the one used for span identification, with the additional Spanish spans.

and 15% of the sentences allocated.

Table 3 presents a high-level statistical characterization of the dataset used in our tests related to span classification. Class disparity is high, with some classes representing less than 0.1% of the dataset. The disparity can be somewhat explained by the type of political rhetoric used in the source materials, where some of the least represented categories are used by a small subset of authors due to their more extreme ideology or controversy. All the assignments rely heavily on the manual encoding performed in the scope of the CAP and MAN projects, and on the annotator's interpretation of the coding instructions[3]. The dataset for span classification is built on the span identification splits, with the addition of the Spanish spans, while attempting to maintain the same category frequency between splits. In total, the train, validation and testing classification splits contain, respectively, 142012, 29736 and 30188 spans.

Data Augmentation Strategies

In some of our tests, we attempted to increase model performance by augmenting the training data, specifically by using Google Translate for translating each span/sentence from the original language into all other languages considered in our study. For each span/sentence S and when considering L different languages, augmentation can easily produce L different spans/sentences (i.e., the original plus the $L-1$ translations). For instance, starting from the 50950 training sentences for span identification in our dataset and as shown in Table 3, the augmentation resulted in 152850 sentences, i.e. a 3 times increase.

Besides translating the complete spans, we also attempted to further augment the dataset for span classification, by considering spans featuring combinations of clauses in different languages, under the intuition that training instances combining different languages can perhaps be beneficial to the training of multilingual models. In particular, the sentences containing adjacent spans of the same class, within the dataset for span identification, can be processed so as to generate additional training spans that combine parts in multiple languages. This technique was only used to augment the training data for the span classification model, given than language changes corresponding to the limits of the spans could easily be picked up when training the model for span identification. The simple translation of training spans roughly triples the size of the dataset into 421601 spans (some spans were not translated as intended and were removed to avoid duplicates). The strategy involving multi-language sentences is able to generate another 54765 new samples, totaling a 335.44% increase over the original training data when paired with the translations.

Automatic Political Content Analysis

This section describes the proposed approaches for span identification and classification within political texts, using contextual language models based on BERT (Devlin et al. 2019), which in turn is based on a multi-layer Transformer (Vaswani et al. 2017) encoder architecture. At its core, a Transformer encoder maps an input sequence of symbol representations into a sequence of continuous represen-

[3]https://manifestoproject.wzb.eu/information/documents/handbooks

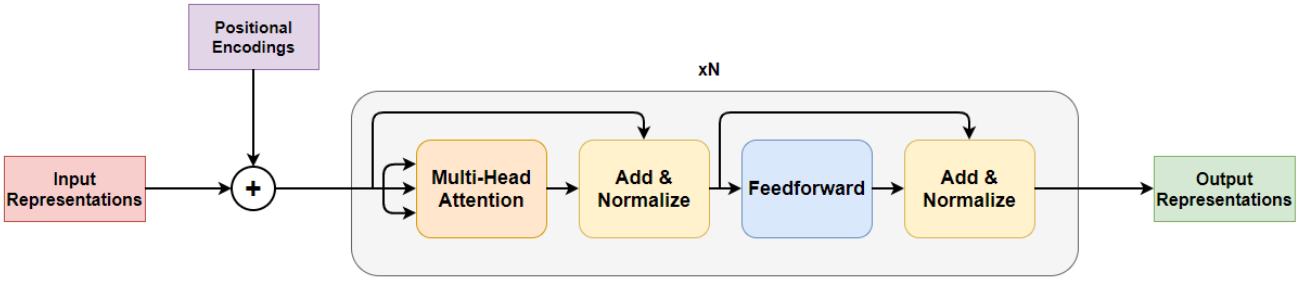

Figure 1: The encoder of the Transformer neural architecture, on which the mBERT and XML-RoBERTa models are based.

tations, and models such as BERT are based on stacking multiple encoder layers. Figure 1 provides an illustration for the encoder of the Transformer architecture, and more details about these models can be obtained in recent surveys (Liu, Kusner, and Blunsom 2020).

In our work, we use pre-trained multilingual models, namely the multilingual BERT (mBERT) and XLM-RoBERTa (Conneau et al. 2019) models. In brief, BERT (Devlin et al. 2019) corresponds to a stack of Transformer encoders pre-trained with two unsupervised tasks, namely a masked language modeling (MLM) task related to predicting masked tokens (i.e., word pieces) from the input, and a next sentence prediction task. The original authors of the BERT model also described the multilingual version, trained on dumps from the top 100 languages with the largest Wikipedias, excluding user and forum pages, with some minor pre-processing and sampling to account for over- and under- represented languages. In turn, XLM-RoBERTa (Conneau et al. 2019) corresponds to a more recent multilingual approach that uses pre-training on a much larger dataset. Beyond the BERT-like MLM pre-training task, this approach also considers a translation language modeling objective (i.e., a MLM task where instead of considering monolingual text, two parallel sentences of different languages are concatenated and, to predict the masked word, the model can either consider the surrounding words of one language or leverage the translated context), which encourages the model to align the multilingual representations.

The fine-tuning of the aforementioned models for both our tasks (i.e., identification and classification) used roughly the same hyperparameters, with the exception of the number of training epochs. These hyperparameters are mostly based on previously published fine-tuning strategies for BERT-based architectures (Sun et al. 2019; Mosbach, Andriushchenko, and Klakow 2020). Training was conducted using an early stopping patience of 2 epochs using the evaluation loss as control metric. Some other key hyperparameters for fine-tuning were an AdamW (Loshchilov and Hutter 2017) epsilon value of 1×10^{-6}, a maximum of 200 tokens in the input sequences, batch sizes of 32 instances, weight decay with $\lambda = 0.01$, and learning rate of 2×10^{-5} linearly increased from 0 in the first 10% of iterations.

We also conduct *in-domain* language model fine-tuning in order to study its potential boost in downstream performance, as seen in Sun et al. (2019). This procedure is accomplished by fine-tuning the language model used for downstream tasks on our training data, using the BERT Masked Language Modeling (MLM) objective (Devlin et al. 2019). This fine-tuning uses the same hyperparameters, with a slightly different batch size of 16 in the XLM-RoBERTa architecture, due to hardware constraints. Approaches with model fine-tuning done prior to the downstream tasks are distinguished in the presentation of results by an -LM suffix.

The best models for both tasks were also used in further experiments considering the augmented datasets mentioned earlier. We then select the best performing model at each individual task to build a complete approach that, given a text of political nature, will identify relevant spans and classify them according to the considered taxonomy.

Relevant Span Segmentation and Classification

The span classification task, assuming previously available spans of text of interest to a particular analysis (e.g. considering a large collection of news headlines), aims to assign one of the possible 46 labels to an entire span. The classification is done using a linear layer on top of the representation produced for the first token by the Transformer language model (i.e., the BERT [CLS] token). As a way to improve the results, the use of a smoothing technique for reference data (*label smoothing*) was also explored. This corresponds to a regularization strategy on top of the cost function (i.e., the categorical cross entropy) that assigns a low ground-truth value, though different from zero, to all MAN classes that belong to the same high-level class (i.e., the same domain) of the category to which the instance is assigned to (Szegedy et al. 2016; Müller, Kornblith, and Hinton 2019).

Large-scale coding initiatives such as the MAN project have noted that political party manifestos are often very long texts, which need to be subdivided into relevant sequences to be classified. Most of these sequences are in the form of entire sentences and, in this case, a full sentence can simply be given a single MAN category. However, this is not always the case, as a sentence can be broken down into a few quasi-sentences, each with its own category.

To identify relevant sequences, we analyze the text at the sentence-level and classify individual tokens through a BIO

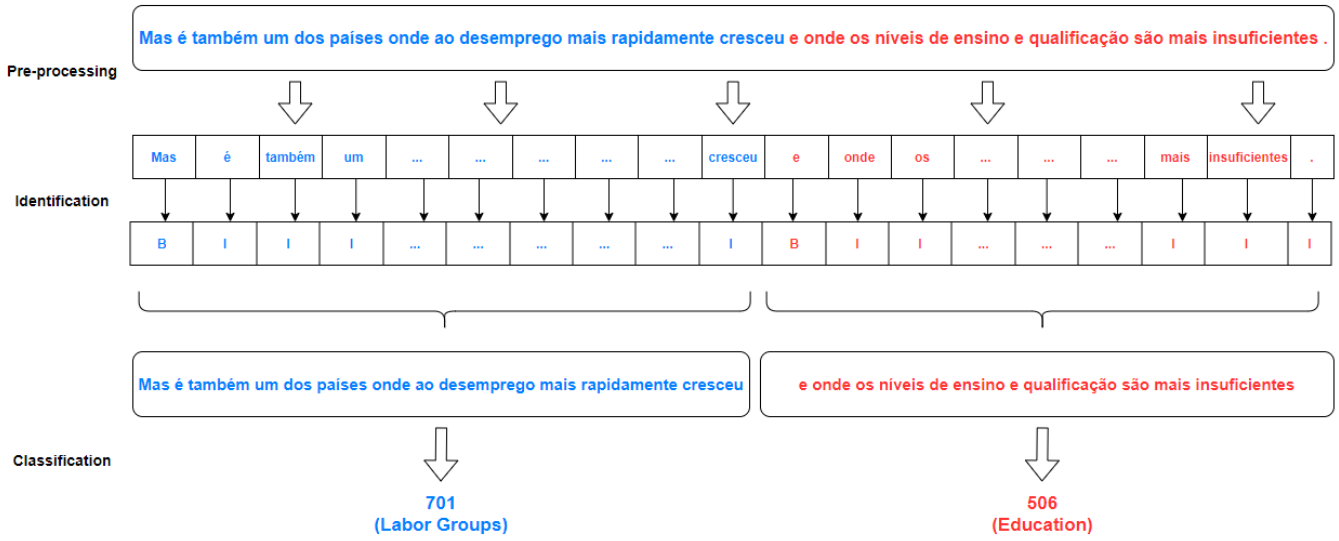

Figure 2: Overview on the complete approach, from input to classification, using a Portuguese multi-span sentence as example.

tagging scheme. Words from spans marked with the General and Headings categories were associated to the outside (O) tag, while the remaining spans were deemed as *relevant*. The first token of a relevant span was identified as beginning (B), while the remaining content was tagged as inside (I). Consistent with the aforementioned quasi-sentence annotations, a sentence can have any number of beginning tags up to the number of quasi-sentences it contains.

While the vast majority of spans in the data collected from the CAP and MAN projects fall inside a 200 token limit (i.e., the limit that, in agreement with the available computational resources, was considered for the input texts in the Transformer models), some outliers still exceed it. The spans that fall outside the limit are processed through an additional step that will attempt to segment them at the nearest comma that would maximize the resulting segments' length. The comma was chosen as delimiter as it is often used as such in the manual processing. In the rare instances where a comma cannot be found, we iteratively remove tokens from the end of the span length until the segment can fit within the limit.

The outputs of the span identification task can be provided as input to the span classifier, in a pipeline approach which involves rebuilding the spans from the BIO annotations and feeding them into the classification model. The complete procedure is illustrated in Figure 2.

Experimental Evaluation

This section describes quantitative experimental results, first discussing the individual performance of the span identification and classification models, before a joint evaluation of the complete pipeline approach.

Results for Relevant Span Identification

While the individual models tended to have similar results at the token-level, with XLM-RoBERTa models slightly outperforming mBERT variants, we decided to also evaluate the models at the span-level, as these results can be more useful for the downstream task of span classification. To that effect, we base our evaluation metrics on those defined for SemEval 2013 Task 9.1 (Segura-Bedmar, Martínez, and Herrero-Zazo 2013), which are based on the MUC metric system (Chinchor and Sundheim 1993). For the evaluation of the span identification model, we chose to focus on the partial and exact metrics, as these only pertain to the span boundaries. The remaining metrics, type and strict, are more appropriate when dealing with the existence of categories associated with the spans, i.e., the complete approach.

The partial metric allows overlaps between relevant spans when considering the evaluation. Partial boundary evaluation considers occurrences of overlaps as having half the value of a complete boundary match. The exact metric, on the other hand, exclusively considers spans whose boundaries were perfectly predicted. Table 4 presents the obtained results, showing that in-domain language model fine-tuning does not increase the performance of the models, with the augmented XLM-RoBERTa being the only one seeing any improvement. Augmentation also seems to be ineffective in this task, having the worst results out of all models. By far, the best model for span identification is the XLM-RoBERTa with no prior in-domain LM fine-tuning, having a superior performance in both partial and exact metrics. As such, we chose this model as the span identification component for our pipeline approach.

Results for Span Classification

Table 5 presents the results for all models, considering the accuracy of classification and the macro-averaged F_1 score as evaluation metrics. As previously mentioned, all models that have been subjected to in-domain language model fine-tuning are identified using the -LM suffix. Initial tests showed that model performance could be increased by ap-

Model	Metric	Overall			Per-Language F_1 Scores		
		Precision	Recall	F_1	Portuguese	Brazilian Portuguese	Italian
mBERT	Exact	0.7835	0.7249	0.7530	0.7621	0.6761	0.8245
	Partial	0.8605	0.7962	0.8271	0.8307	0.7813	0.8779
mBERT-LM	Exact	0.7765	0.7246	0.7497	0.7562	0.6794	0.8237
	Partial	0.8552	0.7980	0.8256	0.8285	0.7832	0.8758
XLM-RoBERTa	Exact	**0.7952**	**0.7289**	**0.7606**	**0.7723**	0.6786	**0.8297**
	Partial	**0.8743**	0.8013	**0.8362**	**0.8438**	0.7849	0.8776
XLM-RoBERTa-LM	Exact	0.7891	0.7288	0.7578	0.7651	**0.6923**	0.8218
	Partial	0.8699	**0.8033**	0.8353	0.8398	**0.7939**	0.8761
XLM-RoBERTa (w/ translations)	Exact	0.7390	0.6970	0.7174	0.7169	0.6536	0.8206
	Partial	0.8405	0.7928	0.8160	0.8140	0.7816	**0.8800**
XLM-RoBERTa-LM (w/ translations)	Exact	0.7453	0.7010	0.7225	0.7245	0.6587	0.8137
	Partial	0.8482	0.7977	0.8222	0.8230	0.7851	0.8770

Table 4: Results for the span identification models.

plying a *label smoothing* (Szegedy et al. 2016) (Müller, Kornblith, and Hinton 2019) technique. While the original implementation (Szegedy et al. 2016) showed improvements, it actually served as inspiration for a slightly different technique that leveraged the hierarchy (i.e., categories and domains) present in the data used. As such, the label smoothing technique, applied to all models in Table 5, assigns 0.9 to the real category and distributes the remaining 0.1 among all the other categories belonging to the same domain. In other words, model training is regularized by minimizing the error of predictions within the correct domain, while still focusing on the correct category. To acquire the results for the domain-level classification one can simply refer to the domain associated to the predicted class, i.e. the first digit of the category. Although the data is imbalanced, the models are still able to accurately classify a majority of spans.

As can be seen in Table 5, the in-domain fine-tuning of the language models, before application in the downstream classification, tends to improve the performance for both mBERT and XLM-RoBERTa, at both levels (i.e., domains and categories). Italian seems to be the worst performing language, perhaps due to it being in the minority when compared with the others (e.g., Portuguese and Brazilian Portuguese are extremely similar, and can more easily influence each other). Spanish has a big gap in its results, but this is most likely due to the non-existence of many rare categories, and the over-representation of many popular ones. To some degree, this is a side-effect of the taxonomy matching.

Data augmentation brought some benefits, although models using this strategy still did not outperform the XLM-RoBERTa-LM model. Prior to the application of our label smoothing strategy, the models using augmented datasets did see a better performance overall. However, these benefits fall short of what can be achieved by the label smoothing regularization technique. Both models using augmented data had similar results, with the translation-only variant having a slight advantage, although the model trained with translations and combinations might not have experienced the full potential benefit from the combination-based augmentations, as these only account for roughly 35% of the original dataset. In light of the results in Table 5, the model chosen for the complete approach's segment classification component was the XLM-RoBERTa-LM trained on the original data, as it showed strong overall performance when compared to the alternatives.

Results for the Complete Approach

In the complete pipeline approach, whose results are reported in Table 6, we use the models with the best individual performance in each task. Therefore, the models for span identification and classification are, respectively, the regular XLM-RoBERTa and the augmented XLM-RoBERTa-LM using span translations. Unlike in the evaluation for span identification, we now consider fine-grained categories assigned to the spans. Therefore, we can use all four sets of metrics from SemEval 2013 Task 9.1, as explained next:

- When span categories are considered, evaluation can be:
  - Typed: only requires a partial overlap between predicted and true span boundaries, but the exact same category is required to be assigned;
  - Strict: requires exact boundary and category matching.
- When not considering span categories, evaluation can be:
  - Exact: requires exact boundary matching, regardless of the matching in the categories;
  - Partial: we require partial boundary matching, regardless of category matching.

The values for the exact and partial metrics are propagated from the identification procedure (Table 4), as these focus solely on the boundaries of the spans, regardless of categories. As expected, the values for the remaining metrics (i.e., type and strict) increase significantly for the coarser level of domains, as was also evident in Table 5. Unlike what happens with the boundary specific metrics, those which consider categories do not have a direct connection with the individual classifier's performance metrics, as they now rely on the output from the span identification component.

When analyzing the results, it is important to not only consider the limitations of the task itself (e.g., ambiguity

Model	Level	Macro-Averaged F_1 Scores					Accuracy
		Portuguese	Brazilian Portuguese	Italian	Spanish	Overall	
mBERT	Categories	0.2916	0.3778	0.2371	0.5336	0.4088	0.6070
	Domains	0.4913	0.6007	0.5018	0.6327	0.6512	0.7042
mBERT-LM	Categories	0.2955	0.3927	0.2367	0.5499	0.4239	0.6117
	Domains	0.4957	0.6112	0.5087	0.6347	0.6561	0.7082
XLM-RoBERTa	Categories	0.3144	0.3966	**0.2672**	**0.5407**	0.4299	0.6111
	Domains	0.5008	0.6230	**0.5103**	0.6279	0.6525	0.7057
XLM-RoBERTa-LM	Categories	0.3123	**0.4162**	0.2662	0.5404	**0.4377**	**0.6202**
	Domains	**0.5075**	0.6220	0.5028	0.6360	0.6588	**0.7138**
XLM-RoBERTa-LM (w/ translations)	Categories	0.3188	0.4143	0.2700	0.5030	0.4321	0.6198
	Domains	0.5009	**0.6253**	0.5011	0.6329	**0.6595**	0.7113
XLM-RoBERTa-LM (w/ translations + combinations)	Categories	**0.3251**	0.3968	0.2511	0.5167	0.4346	0.6175
	Domains	0.4938	0.6100	0.4980	**0.6379**	0.6535	0.7101

Table 5: Results for the span classification models.

Level	Metric	Overall			Per-Language F_1 Scores		
		Precision	Recall	F_1	Portuguese	Brazilian Portuguese	Italian
Categories	Type	0.4688	0.4296	0.4484	0.4524	0.4644	0.4036
	Strict	0.3942	0.3613	0.3771	0.3862	0.3588	0.3600
	Partial	0.8743	0.8013	0.8362	0.8438	0.7849	0.8776
	Exact	0.7952	0.7289	0.7606	0.7723	0.6786	0.8297
Domains	Type	0.6016	0.5514	0.5754	0.5781	0.6013	0.5220
	Strict	0.5030	0.4610	0.4811	0.4891	0.4696	0.4591
	Partial	0.8743	0.8013	0.8362	0.8438	0.7849	0.8776
	Exact	0.7952	0.7289	0.7606	0.7723	0.6786	0.8297

Table 6: Results for the complete approach.

between categories), but also limitations associated to the evaluation procedure in connection with the ground-truth annotations. While we do have a better overview on performance when considering partial overlaps between spans, there are certain situations where we might be penalizing our approach. One example of this is when a series of spans is absorbed into one (i.e., only one span was identified, although there were multiple), which we noticed that occurs frequently. In this situation, if all the real spans had the exact same category, potentially leading to the identification model to not signal a topic shift, only one of the real spans is accounted for in the evaluation metrics. On the reverse situation, where a real span is split into several smaller spans (i.e., over-generation by the identification model, which was relatively rare), all predicted spans may be considered as correct. Naturally, both situations affect the performance metrics, both when categories are considered or not.

Use Case Scenario Related to the Processing of Portuguese Opinion Articles

The multilingual models trained on CAP and MAN data can latter be used to support the analysis of different types of text, including contents associated to online newspapers (e.g., opinion articles, reader comments, etc.) or social media posts. In order to demonstrate a potential use case for our complete approach, we report qualitative results for the processing of opinion articles taken from Portuguese newspapers (i.e., Diário de Notícias, Expresso, Público, Jornal de Notícias, and Observador) during the period between 2015 and 2017. More specifically, we use 942 opinion articles from five nationally recognized authors, namely Alexandre Homem Cristo, Fernanda Câncio, Mariana Mortágua, Rui Ramos, and Rui Tavares. A more detailed characterization of the opinion articles can be found in Table 7.

Using the opinion articles as an example, we start by reporting an analysis of the frequency distribution by class, for the relevant segments recognized in the articles by the complete approach. We also look at the decisions taken by the classification model, using samples taken from the set of opinion articles and using an explanation technique named LIME (Ribeiro, Singh, and Guestrin 2016).

Category Frequency within Opinion Articles

For all authors in the case study, *Political Authority* was the dominant category, ranging between 16% (Mariana Mortágua) and 31% (Alexandre Homem Cristo) of the recognized segments, most likely due to the high volume of comments about the Portuguese Government's performance on various matters. Moreover, several other categories are quite common to all the authors, such as *Democracy* and *Equality*. Figure 3 presents a summary of the results obtained by the proposed approach, considering all the authors.

Alexandre Homem Cristo was parliamentary advisor in

	Descriptive Statistics by Author					
	Alexandre Homem Cristo	Fernanda Câncio	Mariana Mortágua	Rui Ramos	Rui Tavares	Global
Articles	166	34	92	271	379	942
Sentences	5601	787	1925	9579	10060	27952
Words	115847	18124	36584	178043	225616	574214
Average Sentences per Article	34	23	21	35	27	30
Average Words per Article	699	537	398	658	595	610
Average Words per Sentence	21	23	19	19	22	21

Table 7: Characterization of the data set corresponding to the Portuguese opinion articles.

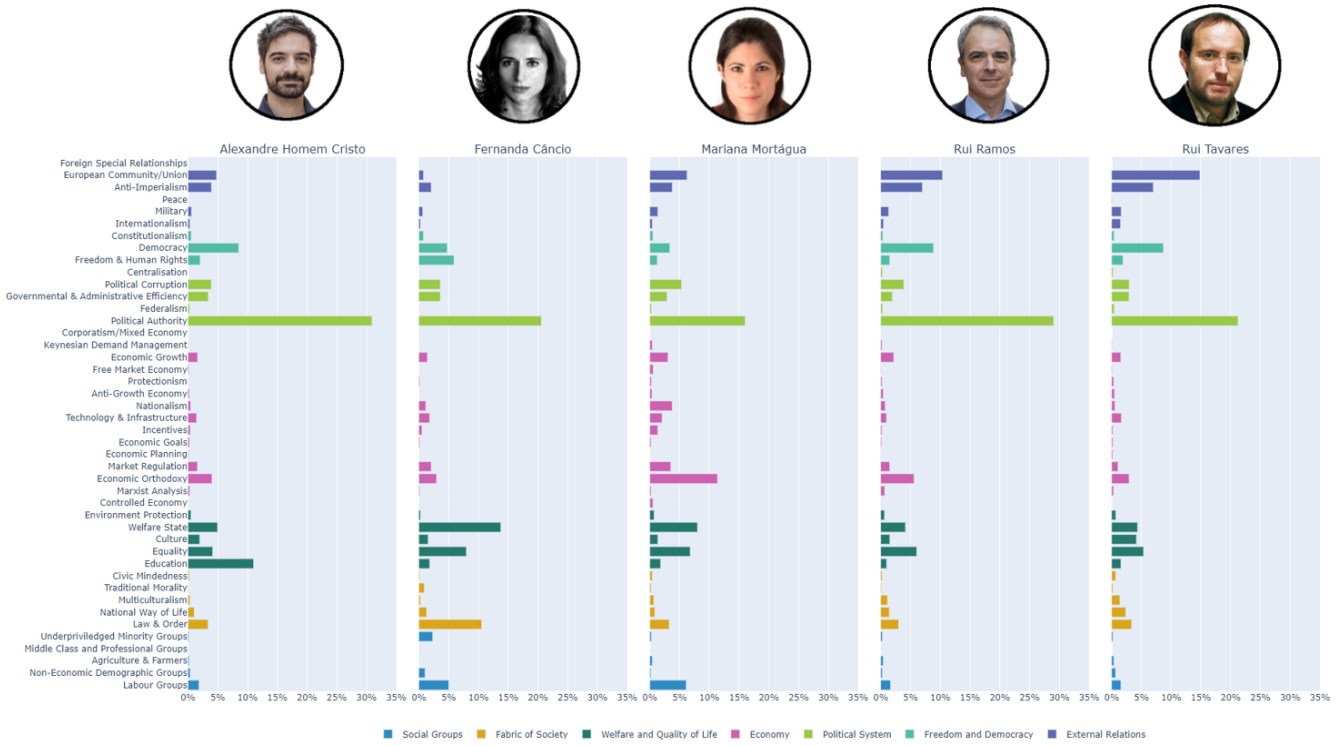

Figure 3: Frequency of the different categories assigned by the model to the opinion articles.

the Assembly of the Republic (2012 – 2015) and counselor in the National Council of Education (2013 – 2015). He writes primarily about *Education*, which justifies that our model observes this category as being the second most present in this author's texts. This author is also the one with the highest percentage of segments marked as belonging to the *Education* (11%) category.

Fernanda Câncio is a well known journalist, tending to address current events. For this reason, it is natural that she is the author with the highest percentage of segments of the *Law & Order* (11%) and *Welfare State* (14%) categories. Among the authors considered, she also stands out in matters like *Labour Groups* and *Freedom & Human Rights*.

Mariana Mortágua is a member of the Assembly of the Republic for the left-wing party Bloco de Esquerda. With a PhD in Economics, she is also the author with the most segments in the domain of Economics, namely with the category *Economic Orthodoxy* being dominant within the considered categories that are related to this topic.

Rui Ramos is co-founder and columnist for the newspaper Observador. He mainly discusses themes related to the political system and external relations, being one of the authors who most discusses themes related to the *European Union* and *Democracy* categories.

Finally, Rui Tavares is a Portuguese historian and politician, and one of the founders of the left-wing party LIVRE. He is the author with the most segments (15%) classified as *European Union*, tending to discuss these issues possibly because he is a former member of the European Parliament.

Interpretability of the Classification Model

Using opinion articles as examples, we can also show the results of attempting to interpret the classifications made by our approach, looking at the text characteristics that may be responsible. It is quite common to categorize *deep learning* models as a black-boxes, where processing is unintelligible. When proposing a model like ours, which aims to categorize political texts, we have to pay special attention not only to

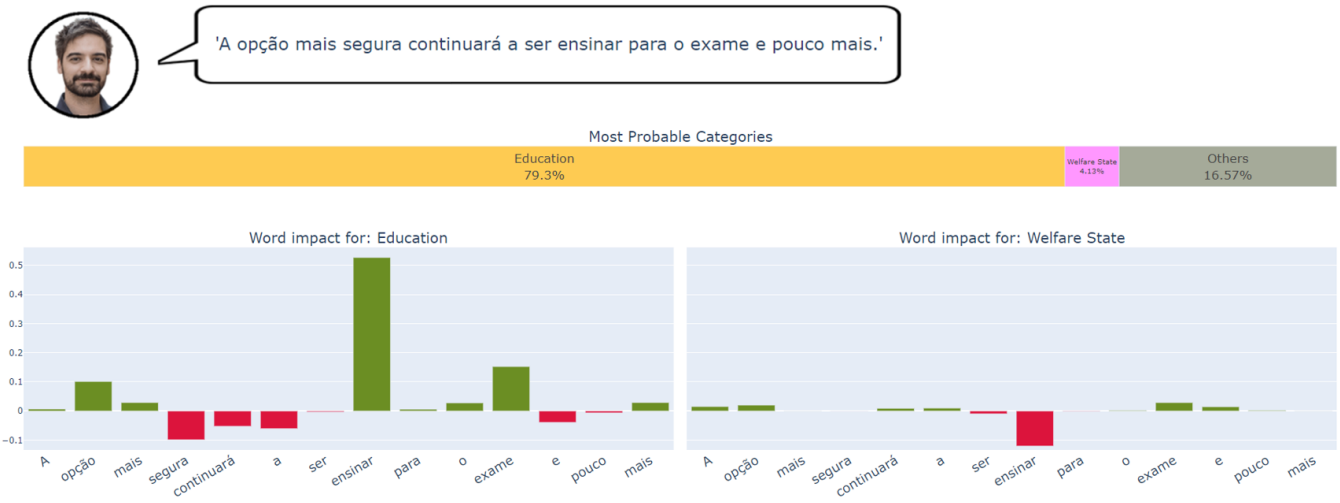

Figure 4: Result interpretation for a segment taken from an opinion article written in May of 2017 by Alexandre Homem Cristo, about access to higher education (*translated: The safest option remains to teach for the exam and little else*).

the final decisions, but also to the rationale behind them, in order to increase confidence in predictions.

In order to address the issue of interpretability, we use a technique that allows us to associate explanations to the decisions made by the model, namely LIME (Ribeiro, Singh, and Guestrin 2016). This technique uses relatively simple models (e.g., linear models defined on an interpretable representation, based on the words of the instances to be analyzed) in order to obtain an approximation of the behavior of the more complex models, from which we extract a readable representation of the rationale of decisions. The LIME technique is independent of the models to which it is applied and, therefore, can be used a posteriori from the training. to explain individual classification decisions. Using LIME, we present an example of the explanations resulting from processing of a text span from the Portuguese opinion articles.

The segment contained in Figure 4, by Alexandre Homem Cristo, was classified by our model as relative to education. As previously observed, this is the author who writes the most about this category. As we can see in the figure, *Education* is the dominant category, with close to 80 percent probability compared to the rest. The excerpt was taken from an article in which the author wrote about access to public universities. Analyzing the results obtained through the LIME tool, we can see the words that weighed most heavily in favor of this category. Specifically, the words *ensinar* (teach) and *exame* (exam), very much related to the subject, have a strong impact. It is possible that the pronounced positive impact of the word *opção* (option) is due to its common use when choosing university preferences, although in this context it does not have that meaning.

Conclusions and Future Work

Manually coding political texts is a difficult, long and monotonous process, often undertaken as a pre-processing step for downstream tasks. Initiatives such as the Manifesto and Comparative Agendas projects have aggregated and evaluated the results from several of these endeavours, from several different countries and languages. These datasets can perhaps now be used to develop automated procedures for content analysis, leveraging recent NLP developments. We specifically implemented a fully automated approach for large-scale analysis of multilingual political texts, leveraging multilingual Transformer-based models.

We presented a taxonomy matching algorithm that was able to map 88% of the categories from the different coding schemes used in the CAP and MAN projects, using their own textual descriptions. The matching of the taxonomies was used to support the creation of datasets combining contents from both these projects, envisioning the training of models that can subsequently be used to analyze different types of textual contents. We then described the use of state-of-the-art multilingual pre-trained language models, fine-tuned with the resulting datasets, in order to identify and classify relevant text spans, emulating the complete manual coding of political texts. Experiments with held-out subsets of the data showed that the models can achieve a high accuracy. Qualitative results were also shown regarding the application to an example downstream task (i.e., the analysis of opinion articles collected from Portuguese online newspapers). We believe the obtained results illustrate the viability of using automated approaches for large scale multilingual studies in the computational social sciences.

In future work, we plan to explore techniques to further improve results in both tasks, such as using different loss functions and more robust model fine-tuning strategies (Jiang et al. 2019), using techniques to better explore the available textual context (Luoma and Pyysalo 2020; Moon et al. 2020; Gunel et al. 2020; Palacio et al. 2021), or using strategies to better deal with highly unbalanced training data (Jiawei et al. 2020; Ye et al. 2020).

References

Ammar, W.; Mulcaire, G.; Tsvetkov, Y.; Lample, G.; Dyer, C.; and Smith, N. A. 2016. Massively multilingual word embeddings. *arXiv:1602.01925* .

Chinchor, N.; and Sundheim, B. 1993. MUC-5 Evaluation Metrics. In *Proceedings of the Message Understanding Conference*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv:1911.02116* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2020. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. *arXiv:2011.01403* .

Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Zhao, T. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv:1911.03437* .

Jiawei, R.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *Proceedings of the Conference on Neural Information Processing Systems*.

Karan, M.; Šnajder, J.; Širinić, D.; and Glavaš, G. 2016. Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts. In *Proceedings of the SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From Word Embeddings to Document Distances. In *Proceedings of the International Conference on Machine Learning*.

Liu, Q.; Kusner, M. J.; and Blunsom, P. 2020. A Survey on Contextual Embeddings. *arXiv:2003.07278* .

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv:1711.05101* .

Luoma, J.; and Pyysalo, S. 2020. Exploring Cross-sentence Contexts for Named Entity Recognition with BERT. In *Proceedings of the International Conference on Computational Linguistics*.

Merz, N.; Regel, S.; and Lewandowski, J. 2016. The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics* 3(2).

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Conference on Neural Information Processing Systems*.

Moon, S. J.; Mo, S.; Lee, K.; Lee, J.; and Shin, J. 2020. MASKER: Masked Keyword Regularization for Reliable Text Classification. *arXiv:2012.09392* .

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv:2006.04884* .

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In *Proceedings of the Conference on Neural Information Processing Systems*.

Palacio, S.; Engler, P.; Hees, J.; and Dengel, A. 2021. Contextual Classification Using Self-Supervised Auxiliary Models for Deep Neural Networks. *arXiv:2101.03057* .

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Ruder, S.; Vulić, I.; and Søgaard, A. 2019. A Survey of Cross-Lingual Word Embedding Models. *Journal of Artificial Intelligence Research* 65.

Segura-Bedmar, I.; Martínez, P.; and Herrero-Zazo, M. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Proceedings of the International Workshop on Semantic Evaluation*.

Subramanian, S.; Cohn, T.; and Baldwin, T. 2018. Hierarchical Structured Model for Fine-to-Coarse Manifesto Text Analysis. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Subramanian, S.; Cohn, T.; Baldwin, T.; and Brooke, J. 2017. Joint Sentence-Document Model for Manifesto Text Analysis. In *Proceedings of the Australasian Language Technology Association Workshop*.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2019. How to fine-tune BERT for text classification? In *Proceedings of the National Conference on Chinese Computational Linguistics*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the Conference on Neural Information Processing Systems*.

Ye, H.-J.; Chen, H.-Y.; Zhan, D.-C.; and Chao, W.-L. 2020. Identifying and Compensating for Feature Deviation in Imbalanced Deep Learning. *arXiv:2001.01385* .