Joint Sentence-Document Model for Manifesto Text Analysis

Anonymous ALTA submission

Abstract

Political parties are essential institutions of democracy. Election programs (so-called manifestos) is a published verbal declaration of the intentions, motives, or views of the political party.[1] Political scientists use manifestos to understand sentence level policy relevant themes discussed and also quantify manifesto's position on the left-right spectrum. Rather than handling the two tasks separately, we propose a joint sentence-document model for sentence level thematic classification and document level position quantification using manifestos in different languages. Inorder to handle text from multiple languages we exploit continuous neural embeddings for semantic text representation. We have empirically shown the effectiveness of proposed approach using manifestos from thirteen countries which are written in six different languages.

1 Introduction

Political parties are at the core of contemporary democratic systems. One of the widely used dataset by political scientists includes the Comparative Manifesto Project (CMP) dataset, initiated by (**?**), that collects party election programs (so-called manifestos) from elections in many countries around the world. The goal of the project is to provide a large data collection to support political studies on electoral processes. A sub-part of the manifestos has been manually annotated at sentence-level with one of over fifty fine-grained political themes, divided into 7 coarse-grained topics.[2] Each sentence is segmented (if it discusses more than one topic) and labeled. While manual annotations are very useful for political analyses, they come with two major drawbacks. First, it is very time-consuming and labor-intensive to manually annotate each sentence with the correct category from a complex annotation scheme. Secondly, coders' preferences towards particular categories might cause annotation inconsistencies and affect comparability between manifestos annotated by different coders (**?**). (**?**) is one of the early works using manifesto text for automatic thematic classification, using a topic hierarchy defined by Dutch political experts. In (**?**) authors worked with CMP dataset to do coarse-level topic classification (defined by CMP). They used Markov Logic Networks (MLN) to model sentence adjacency topic smoothness constraint. Nonetheless, manually coded manifestos remain the crucial data source for studies in computational political science (**??**).

Existing classification models, except (**?**), utilize discrete representation of text (i.e., bag of words) and can thus exploit only monolingual data (i.e., train and predict same language instances). Hence, most of the works analyze manifesto text at country level. Recent work has shown the use of neural embeddings for multi-lingual manifesto text analysis (**?**). In (**?**) authors show that multi-lingual embeddings are more effective for cross-lingual coarse level manifesto text topic classification where there is labeled data only in source language. They have also shown that for mono-lingual setting, with sufficient labeled data in each language, bag-of-words representation is more effective than using continuous representation.

Other than the sentence level labels, the manifesto text also has document level signals such as

[1] https://en.wikipedia.org/wiki/Manifesto

[2] https://manifestoproject.wzb.eu/coding_schemes/mp_v5

RILE score — defined as the difference between count of sentences discussing left and right topics (formulation is given below), expert left/right survey scores and voter party position survey scores (?).

$$RILE = \sum_{r \in R} per_r - \sum_{l \in L} per_l$$

where, R = {104, 201, 203, 305, 401, 402, 407, 414, 505, 601, 603, 605, 606} and L = {103, 105, 106, 107, 202, 403, 404, 406, 412, 413, 504, 506, 701}, "per_{xyz}" denotes share of each topic (xyz) as given in Figure 2, per document. Almost all the works use label count aggregation of manually annotated sentences as features (??) and not text of the document for quantification task. In this work we evaluate the use of embeddings for multi-lingual manifesto text document level position quantification based on RILE scores. Secondly, works till now handle sentence level classification and document level quantification tasks separately. Rather than handling it separately, we propose a joint sentence-document model to handle both the tasks together.

2 Related Works

The recent adoption of NLP methods had led to significant advances in the field of Computational Social Science (CSS) (?) and political science in particular (?). Among other tasks, researchers have addressed the identification of political differences from text (??), positioning of political entities on a leftright spectrum (?), as well as the detection of political events (Nanni et al., 2017) and prominent topics (Lauscher et al., 2016) in political texts. For what concerns the analysis of manifestos previous studies have focused on topical segmentation (Glavas et al. , 2016) and monolingual (English) classification of sentences into coarse-grained topics (Zirn et al., 2016). Because manifesto sentences are short and short text classification is inherently challenging due to limited context, Zirn et al. (2016) proposed to apply a global optimization step (performed via Markov Logic network) on top of independent topic decisions for sentences. Numerous supervised models have also been proposed for classification of other types of political text (Purpura and Hillard, 2006; Stewart and Zhukov, 2009; Verberne et al., 2014; Karan et al., 2016, inter alia). However, these models also represent texts as sets of discrete words which directly limits their applicability to monolingual classification settings only.

3 Appendix

Domain 1: External Relations
101 Foreign Special Relationships: Positive
102 Foreign Special Relationships: Negative
103 Anti-Imperialism
104 Military: Positive
105 Military: Negative
106 Peace
107 Internationalism: Positive
108 European Community/Union: Positive
109 Internationalism: Negative
110 European Community/Union: Negative
Domain 2: Freedom and Democracy
201 Freedom and Human Rights
202 Democracy
203 Constitutionalism: Positive
204 Constitutionalism: Negative
Domain 3: Political System
301 Decentralisation
302 Centralisation
303 Governmental and Administrative Efficiency
304 Political Corruption
305 Political Authority
Domain 4: Economy
401 Free Market Economy
402 Incentives: Positive
403 Market Regulation
404 Economic Planning
405 Corporatism/Mixed Economy
406 Protectionism: Positive
407 Protectionism: Negative
408 Economic Goals
409 Keynesian Demand Management
410 Economic Growth: Positive
411 Technology and Infrastructure: Positive
412 Controlled Economy
413 Nationalisation
414 Economic Orthodoxy
415 Marxist Analysis
416 Anti-Growth Economy: Positive
Domain 5: Welfare and Quality of Life
501 Environmental Protection
502 Culture: Positive
503 Equality: Positive
504 Welfare State Expansion
505 Welfare State Limitation
506 Education Expansion
507 Education Limitation
Domain 6: Fabric of Society
601 National Way of Life: Positive
602 National Way of Life: Negative
603 Traditional Morality: Positive
604 Traditional Morality: Negative
605 Law and Order: Positive
606 Civic Mindedness: Positive
607 Multiculturalism: Positive
608 Multiculturalism: Negative
Domain 7: Social Groups
701 Labour Groups: Positive
702 Labour Groups: Negative
703 Agriculture and Farmers: Positive
704 Middle Class and Professional Groups
705 Underprivileged Minority Groups
706 Non-economic Demographic Groups

000 No meaningful category applies

Figure 1: CMP coding scheme (taken from coding instructions). *Left* topics are given in red, *right* topics are given in blue and the rest are given in black