“We will Reduce Taxes” – Identifying Election Pledges with Language Models

Tommaso Fornaciari
Università Bocconi
fornaciari@unibocconi.it

Dirk Hovy
Università Bocconi
dirk.hovy@unibocconi.it

Elin Naurin
University of Gothenburg
elin.naurin@pol.gu.se

Julia Runeson
University of Gothenburg
julia.runeson@gu.se

Robert Thomson
Monash University
robert.thomson@monash.edu

Pankaj Adhikari
Monash University
pankaj.adhikari@monash.edu

Abstract

In an election campaign, political parties pledge to implement various projects–should they be elected. But do they follow through? To track election pledges from parties' election manifestos, we need to distinguish between pledges and general statements. In this paper, we use election manifestos of Swedish and Indian political parties to learn neural models that distinguish actual pledges from generic political positions. Since pledges might vary by election year and party, we implement a Multi-Task Learning (MTL) setup, predicting election year and manifesto's party as auxiliary tasks. Pledges can also span several sentences, so we use hierarchical models that incorporate contextual information. Lastly, we evaluate the models in a Zero-Shot Learning (ZSL) framework across countries and languages. Our results indicate that year and party have predictive power even in ZSL, while context introduces some noise. We finally discuss the linguistic features of pledges.

1 Introduction

Before the election, political parties summarize what they pledge to the voters in manifestos published for anyone to read. These are often seen as an essential document for the representatives to refer to when explaining to the voters what they wish to do, should they gain their votes.

However, there is a difference between pledging and fulfilling. Political scientists are highly interested in whether pledges were fulfilled, a question that is gaining a growing interest in the broader scientific community (Naurin et al., 2019). Several approaches exist, but they are primarily confined to manual analysis of individual countries or elections. They indicate that governmental parties mostly fulfill their election pledges (Naurin et al., 2019; Thomson et al., 2017), but there are too many elections worldwide to analyze all campaign pledges manually. We need automated ways to identify pledges and hold governments accountable systematically.

Checking whether a pledge was fulfilled still requires manual work by trained political scientists, but the first step–*identifying* pledges–is a problem very much made for NLP. First, NLP can automate pledge identification to distinguish pledges from irrelevant content. This allows the study of pledge fulfillment at scale. An average election manifesto in our corpus has 418 sentences, but only 118 of them (27.5%) will contain a pledge. The rest is filler material. It takes several days to train an annotator, who then spends around 6-8 hours on a single manifesto, to identify those 27.5% of pledges. Cutting down on this laborious first part frees up time to focus on the more complex issue of determining whether those pledges were fulfilled. Second, NLP methods can help us to understand the linguistic style and communication strategies associated with election pledges. This interpretation is necessary for social sciences to understand how political messages are structured and conveyed.

This paper presents neural pledge identification models to address these two points. We use a data set of almost 13k sentences from election manifestos concerning the last 25 years and 11 parties from Sweden and India. Each sentence is annotated as including a pledge (“pledge”) or not including a pledge (“non-pledge”). We implement several deep neural models based on BERT (Devlin et al., 2018). We use its Swedish, English (for the Indian data), and multi-lingual (mBERT) versions. We feed the BERT's output into customized attention mechanisms to detect specific pledge-related patterns. We compare our neural models with a Logistic Regression baseline that the deep models easily outperform.

However, pledges could not just depend on some

Corpus	Text	Class
Swedish	Vi i Centerpartiet är stolta över vad vi uppnått i regeringen.	non-pledge
	In the Center Party we are proud of what we achieved in the government.	
	Barnkonventionen ska göras till svensk lag.	pledge
	The Convention on the Rights of the Child shall be made Swedish law.	
Indian	They have neither competence nor commitment.	non-pledge
	Five new IITs will be established before 2005.	pledge

Table 1: Examples of pledges and non-pledges from Swedish and Indian manifestos.

signal words or expressions. References to the environment might be core pledges for one party, but just commentary for another. Specific issues will be pledge-worthy one year (think pandemic responses), but not in others. To measure the effects of all of these confounds (i.e., *election year* and *party*), we adopt a Multi-Task Learning (MTL) framework. The main task is to classify sentences as pledge or non-pledge, with auxiliary tasks predicting the year, party, or both. We identify the conditions where MTL models with year and party improve the models' performance, indicating when these two factors are useful confounds. There seem to be stark differences between countries, though: even using a multi-lingual approach (which has access to more training data) does not improve on language-specific approaches.

We are also interested in zero-shot learning, i.e., training models on data from a Country and testing it on a different country. This allows evaluating the possibility of working on pledges directly, without any previous manual annotation. It turns out that the models perform reasonably well despite the challenging conditions. However, the differences between test Countries indicate that pledges are not as universal as we might think.

Somewhat surprisingly, we also find that incorporating the context of any sort (that is, one or more sentences preceding the target text) does not help but hurts performance. This is presumably because pledges are rare, and context introduces more noise than signal. Our data and our models are available at `www.anonymized.com`.

This work is part of a larger interdisciplinary project within which we are also interested in learning more about pledges' nature. I.e., what their linguistic features and patterns are. To gain those insights, we extract the Information Gain value (Forman, 2003) of 1–4-grams and visualize the model's decisions via the Sampling and Occlusion (SOC) algorithm (Jin et al., 2019). SOC provides a hierarchical view of BERT's most informative linguistic patterns in the classification.

Contributions The contributions of this paper are: 1) We provide a new, multi-lingual corpus of election manifestos from Swedish and Indian parties, annotated at sentence level as pledges or non-pledges; 2) We are the first to apply neural models to the task of election pledge classification, accounting for confounds; 3) We provide insights about the linguistic features of election pledges and the models' interpretation.

2 Data

We collect and annotate a corpus of election manifestos from two countries: Sweden and India. The texts are in Swedish and English, respectively. We provide some examples in Table 1.

The Swedish data contain 5098 instances from 9 parties and 6 elections, ranging from 1994 to 2014. The rate of pledges per manifesto is 32.09%. These texts are also part of the corpus of the Manifesto Project (MP) (Volkens et al., 2012; Merz et al., 2016, Section 7), but in our case, we also follow a different, new annotation scheme. Following the Comparative Party Pledges Project (CPPP) of Naurin et al. (2019) and Thomson et al. (2017), we further distinguish between broad and narrow pledges, i.e., between generic and detailed commitments to undertake determined actions. We have 23.32% narrow and 8.77% broad pledges in the Swedish data.

The Indian texts contain 7729 sentences from 2 parties and 5 election cycles, from 1999 to 2019.[1] Here, the annotators only distinguished sentences including a narrow pledge from non-pledge sentences, with a pledge rate of 24.52%.

In total, we have 12827 sentences and 3531

[1] If referring to the data set, we will use *Indian*, but if referring to the language, we use *English*.

pledges (27.53%). Since we only have binary labels for the Indian data, we combine broad and narrow pledges in the Swedish corpus.[2]

2.1 Annotation process

According to CPPP, an election pledge is a statement that can be tested for fulfillment. In practice, this means that the annotator requires contextual knowledge of the country and its national politics to decide. We trained Swedish and Indian annotators to label the Swedish and Indian manifestos for our study, respectively. Two domain experts conducted the training, one for each data set. The two annotators communicated with the two domain experts throughout the annotation process, to handle complicated cases. Hence, four people were involved in the annotation of the manifestos.

Moreover, to test agreement in the Indian data, three trained annotators labeled 100 sentences. Their Krippendorff's α and Fleiss's κ are 0.65. On the Swedish data, two trained annotators labeled 100 sentences again, with Krippendorff's α and Cohen's κ at 0.61. In both cases, the agreement can be considered as 'substantial' (Landis and Koch, 1977). Our results are coherent with that reported by Naurin et al. (2019).

3 Methods

We have three experimental conditions: 1) Swedish texts alone, 2) Indian texts alone, and 3) Swedish and Indian texts together (multilingual condition). The last condition allows us to see whether performance improves with access to more training data and whether pledges are comparable across countries.

As a baseline, we train Logistic Regression models, optimized with the parameter $C = 1$, based on TF-IDF-weighted Bag-Of-Words (BOW) from $1-$ to $3-$grams, with document frequency range from 0.001 and 0.75. Table 2 show the performance. We evaluate our models with standard metrics: precision, recall, and F1-measure averaged between the two classes.

3.1 Neural models

For the first two experimental conditions, we consider separate, mono-lingual Swedish and English BERT models and the multi-lingual (mBERT) version. In the third experimental condition, where we merge the two data sets, we can only use the multi-lingual BERT.

	Acc.	Prec.	Rec.	F1
Sweden	76.30	73.53	75.73	74.17
India	77.90	71.81	76.15	73.10
Both	77.28	72.88	76.57	73.90

Table 2: Logistic Regression baselines.

Single-Task Learning. Our base models are binary classifiers, i.e., single-tasks (STL) models. Standard BERT classifiers perform the task with a fully connected layer on top of the BERT's output. In contrast, we reframe the BERT's `[CLS]` token representation as a single-row matrix, and we feed it into a single-layer, single-head Transformer (Vaswani et al., 2017). In our pilots, we found that this specialized structure allows us to detect specific pledge's patterns from the BERT representation more effectively than a standard output dense layer alone. Finally, the Transformer is connected to a dense output layer for the prediction.

Multi-Task Learning. We implement three different MTL versions, differing by the auxiliary task combinations. We have two potential auxiliary tasks: predicting the *election year* and the *party* that produced the manifesto. We add a further dense output layer to the base model to perform the MTL tasks: 1) predicting the election year, 2) the party, or 3) both. We use the mean of the task losses for error-backpropagation in the MTL networks. Since their magnitude is bounded by the fact that all predictions are probability distributions, no normalization is needed. Figure 1 (left) shows the scheme of the MTL models.

Contextual models. We also build models considering the sentence preceding the target text as context, allowing us to test its impact on classification performance. We incorporate the context sentence in two state-of-the-art ways: through pair-BERT, which accepts two texts as input, and through a hierarchical model. In the first case, the model is structurally equivalent to the base model: only the input representation for BERT changes to include two sentences, separated by the separator token `[SEP]`. In the second case, we stack the representations of the BERT classification tokens (`[CLS]`) of both context and target sentences and

[2] We trained binary classifiers for narrow pledges in a pilot study, treating broad pledges as non-pledges. The performance was slightly worse than in the case reported here, due to a more noisy "non-pledge" class and more skewed class balance. We include those results in Appendix.

BERT	Task	Target	acc	prec	rec	f1
Swedish	STL	Sweden	87.01	85.19	84.87	85.03
Swedish	MTL Party	Sweden	87.11	85.35	84.89	85.12
Swedish	MTL Year	Sweden	87.07	85.37	84.74	85.04
Swedish	MTL Party + Year	Sweden	87.05	85.45	84.51	84.95
Multilingual	STL	Sweden	81.94	79.57	78.20	78.81
Multilingual	MTL Party	Sweden	81.78	79.35	78.10	78.66
Multilingual	MTL Year	Sweden	82.05	79.77	78.14	78.85
Multilingual	MTL Party + Year	Sweden	81.83	79.55	77.80	78.54
Multilingual	STL	India (0-shot)	73.15	68.78	74.34	69.27
Multilingual	MTL Party	India (0-shot)	**75.33** **	**69.24** *	73.45	**70.31** **
Multilingual	MTL Year	India (0-shot)	**74.4** **	69.08	74.05	**69.96** **
Multilingual	MTL Party + Year	India (0-shot)	**75.67** **	**69.17** *	72.83	**70.25** **

Table 3: Training data set: Sweden. Language: Swedish. Significance of MTL over STL: $^{**}: p \leq 0.01$; $^{*}: p \leq 0.05$

feed them into a Transformer connected to a dense layer that gives the output. Figure 1 (right) depicts its structure.

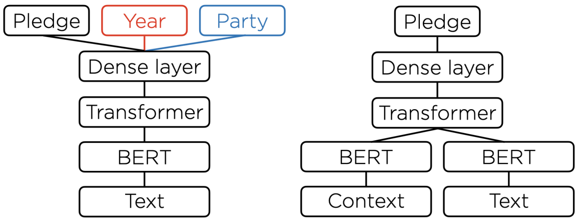

Figure 1: Left: STL and MTL model scheme. STL: black boxes. MTL with one auxiliary task: black + red or black + blue boxes. MTL with two auxiliary tasks: all boxes. Right: Hierarchical models' scheme

Settings and significance tests. To reduce the variability of the models' random initialization and to make our results more robust, we run 10 repeats for each experimental condition and compute the overall performance. To test the significance of the improvements over the base model, we use a bootstrap sampling test on all runs (Søgaard et al., 2014), with 1000 loops and a sample size of 30%.

For each experiment, we run 10-fold cross-validation. In each fold, we use 80% of texts as training set, 10% for development and 10% for test. In the ZSL experiments, we use 90% and 10% of a data set for training and development, respectively, and the whole other data set as test set.

For the main task, the loss function is the binary (sigmoid) cross-entropy; it is (soft-max) cross-entropy for the auxiliary tasks. We use the Adam optimizer (Kingma and Ba, 2014). We select the models through an early-stopping that requires a decrement rate on the development set's loss lower than 8% for five consecutive epochs. Our learning rate is 0.002, drop-out probability 0.3, and batch size 512, manually tuned.

4 Experiments

We report results on all models for each of the three experimental conditions: 1) Swedish corpus encoded with Swedish and multi-lingual BERT (Table 3); 2) Indian corpus encoded with English and multi-lingual BERT (Table 4); and 3) the joint Swedish and Indian data together, encoded with multi-lingual BERT (Table 5).

For each of these conditions, we train a baseline Logistic Regression model (Section 3)—an STL base model as described in Section 3.1—and compare them with MTL and contextual models. Since all the models outperform the Logistic Regression baseline, we report significance levels concerning the improvement over the STL models.

5 Results

We see a substantial performance difference between the two BERT encodings (Swedish and mBERT) regarding the Swedish data. The Swedish version outperforms the multi-lingual one and reaches the best performance of the experiments' set (Table 3).

We do not see the same performance difference in the Indian data, where English and multi-lingual BERT produce similar outcomes, with the multi-lingual even slightly better. Results are generally

BERT	Task	Target	acc	prec	rec	f1
English	STL	India	83.74	78.98	74.57	76.31
English	MTL Party	India	83.63	78.71	74.64	76.27
English	MTL Year	India	83.91	**79.62** *	74.1	76.16
English	MTL Party + Year	India	83.89	78.84	**75.68** **	**77.02** *
Multilingual	STL	India	83.49	78.08	75.64	76.71
Multilingual	MTL Party	India	83.48	78.41	74.58	76.14
Multilingual	MTL Year	India	83.70	**78.81** *	74.78	76.41
Multilingual	MTL Party + Year	India	83.69	**78.75** **	74.84	76.42
Multilingual	STL	Sweden (0-shot)	73.57	75.32	60.73	60.44
Multilingual	MTL Party	Sweden (0-shot)	72.29	76.19	57.95	56.05
Multilingual	MTL Year	Sweden (0-shot)	72.28	**76.32** *	57.91	55.98
Multilingual	MTL Party + Year	Sweden (0-shot)	72.69	**76.58** *	58.63	57.12

Table 4: Training data set: India. Language: English. Significance of MTL over STL: $^{**}: p \leq 0.01$; $^{*}: p \leq 0.05$

BERT	Task	Target	acc	prec	rec	f1
Multilingual	STL	Both	82.74	78.70	76.84	77.67
Multilingual	MTL Party	Both	82.80	79.05	76.17	77.37
Multilingual	MTL Year	Both	82.53	78.48	76.42	77.32
Multilingual	MTL Party + Year	Both	82.73	78.91	76.19	77.33

Table 5: Data set: Sweden & India. Language: Swedish and English.

lower than those for the Swedish data (Table 4).

As expected, the results of the multi-lingual model trained on the joint data set lie between the respective multi-lingual models on the two data sets separately. So while the Swedish BERT is clearly more effective than the multi-lingual one on Swedish texts, the amount of data in the multi-lingual language model presumably counteracts the lack of annotated data in the Indian data set.

MTL vs STL. The MTL models are effective in some cases.First, it helps in the ZSL conditions. This suggests that training the models to contextualize the notion of pledge with respect to party and year reduces overfitting. Also, when effective, MTL models improve the precision. This is an expected effect, as the models learn to detect pledges as well as historical periods and political areas. This is an interesting feature for ZSL, where the confidence regarding the positive cases' identification is more valuable than a good recall. We also tested the MTL models in case of reduced amount of data. In particular, we trained models considering the election manifestos from 2000 only. We found that the MTL give a stronger contribution in such conditions. The results of these experiments are included in Appendix.

Does Context Help? In a word, no. While a disappointing outcome, we find it important to include this finding here, as it goes very much against both intuition and prior research. Bilbao-Jayo and Almeida (2018), for example, found that contextual information is helpful when classifying political topics (see Section 7). Election pledges seem to be more self-contained statements, relying on linguistic formulas that make them recognizable (and probably memorizable) regardless of their linguistic context (Section 6).

We explored two different models to incorporate the sentence preceding the target texts. In both cases, though, we consistently find that the previous sentence's contextual information adds more noise than a useful signal for prediction. The fall is from moderate to drastic (up to 10 points in F1), particularly for the pair-BERT models where, by design, target and context representations are not trainable. The hierarchical models' performance is more stable, but the context does not improve the performance.

	IG	Fr.		IG	Fr.
Vi_vill_också	0.013561	21	Alliansen_har_följande	0.005713	26
Ett_införande_av	0.008383	13	har_följande_skarpa	0.004391	20
•_Ett_införande	0.008383	13	följande_skarpa_förslag	0.004391	20
•_Ett_utökat	0.005799	9	I_vårt_Sverige	0.004205	41
Ett_utökat_stöd	0.004509	7	vill_under_kommande	0.003072	14
•_En_satsning	0.004509	7	ska_vara_ett	0.003072	14
•_En_utökad	0.004509	7	Alliansen_vill_under	0.003072	14
utökat_stöd_till	0.004509	7	under_kommande_mandatperiod	0.003072	14
utökad_satsning_på	0.004509	7	kommande_mandatperiod_att:	0.003072	14
så_att_det	0.004177	10	Det_är_en	0.002632	12

Table 6: Swedish tri-grams indicative of pledge (left) and non-pledge (right)

	IG	Fr.		IG	Fr.
will_be_set	0.020450	49	It_is_the	0.006252	40
be_set_up	0.017386	41	the_Congress_that	0.005938	38
will_be_launched	0.015698	26	is_the_Congress	0.005311	34
will_set_up	0.014979	20	National_Congress_will	0.004462	95
the_next_five	0.013200	23	will_be_made	0.003918	63
in_the_next	0.011840	19	The_Congress_will	0.003897	60
set_up_a	0.011294	27	the_Congress_is	0.003744	24
in_five_years.	0.010608	15	A_time_to	0.003431	22
be_launched_to	0.010163	17	has_always_been	0.003275	21
over_the_next	0.009743	14	It_is_a	0.002962	19

Table 7: Indian tri-grams indicative of pledge (left) and non-pledge (right)

6 The language of pledges

To better understand the pledges' linguistic features, we follow two strategies: 1) computing the Information Gain (IG) of word n-grams, and 2) using the Sampling and Occlusion (SOC) algorithm (Jin et al., 2019).

Information Gain measures the entropy of (sequences of) terms between the different classes. The more imbalanced the presence of such terms for one label class at the other's expense, the higher the IG value. Tables 6 and 7 show the trigrams with the highest IG values (and relative frequencies), divided according to the class of which they are indicative, i.e., where they are more frequently found. While we computed the IG score from uni-grams to $penta$-grams, we show only tri-grams that, for illustration, represent the best trade-off between meaningful and frequent chunks of text. For the full translation of the Swedish texts, see Appendix.

These n-grams suggest that a formulaic language characterizes election pledges: stereotypical expressions characterize specific sentences as pledges. For example, in the Swedish data set, the bullet is used as a clear marker that introduces statements of some form of commitment. We also find expressions indicating volition ("Vi **vill** också" – "We also **want**..."), consequences ("**så** att det" – "**So** that..."), future ("**will** be set", "**will** be launched") and determined temporal horizons ("in five years", ”over the next”). In contrast, both in the Indian and the Swedish data, references to political entities such as parties ("Alliansen"), congresses (”National Congress”) and even countries ("Sverige", "India") are associated with non-pledge texts: they refer, more probably, to broad political positions or to claims about the past ("has always been", "ska vara ett" – "should be one").

Interestingly, the phrase "skarpa förslag" does *not* signal pledges, even though it means "specific policy proposals" (which are essentially the same as pledges). This distinction indicates that this phrase merely introduces pledges or provides a strong language for un-testable policy statements (such as "we promise safety to all children" or

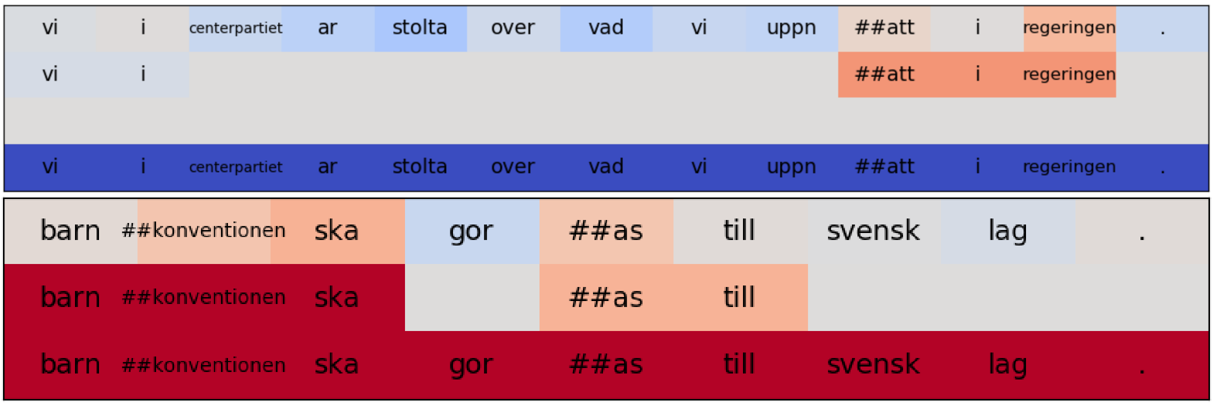

Figure 2: Output of the SOC algorithm on the Swedish corpus. The red terms predict Pledge, the blue ones predict Non-pledge.

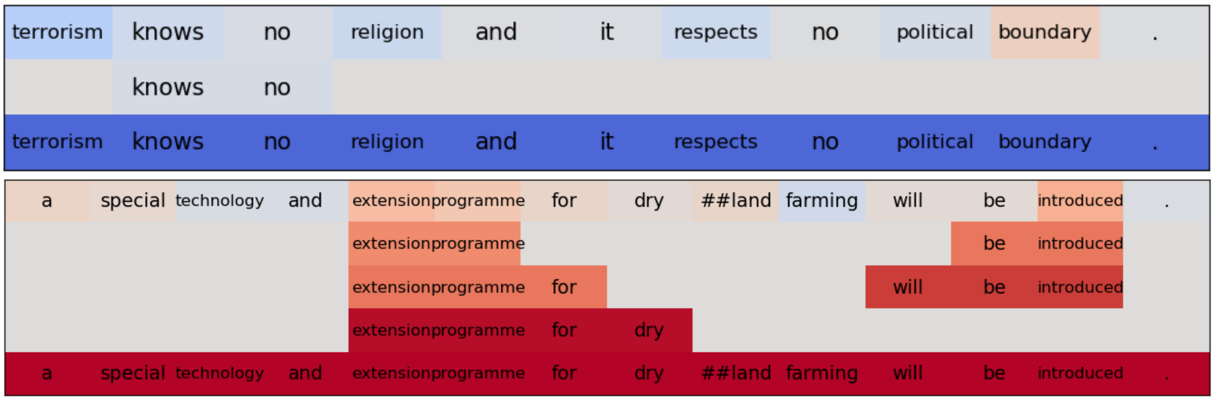

Figure 3: Output of the SOC algorithm on the English corpus. The red terms predict Pledge, the blue ones predict Non-pledge.

"we will put forward strict legislation to make our country safe again").

Given the relatively limited frequency of the selected n-grams, we did not measure the IG stratification by party and/or election year. However, given the relative MTL models' success, we hypothesize that, with more data, it will be possible to identify specific trends for political areas and historical moments.

Aware that the patterns detected by the neural models are not necessarily interpretable in terms of human common sense, we also wanted to highlight the words that the models find to be the most influential for their output. These patterns can feedback into the interpretation of pledge structures and mechanisms by social scientists.

We also use the Sampling and Occlusion (SOC) algorithm (Jin et al., 2019), a *post-hoc* explanation algorithm that measures the importance of specific words in a sentence by considering the prediction difference after replacing each word with a MASK token (Jin et al., 2019). Since the outcomes depend on the context words, but Jin et al. (2019) are interested in the single words' relevance, they do not use the whole context but sample words from it. In this way, they reduce the context's weight, emphasizing that of the word itself.

Figure 2 and 3 show four examples of correctly classified sentences, two pledges and two non-pledges from Swedish and English language respectively (the same as shown in Table 1). The model interprets the red words as indicative of pledges, the blue ones of non-pledges. However, they cannot be interpreted as representative of the overall models' functioning. Even so, they show how generic words such as "stolta" ("proud") are indicative of non-pledges, while expressions indicating commitment ("ska göras till" – "to be made to") and concrete topics ("Barnkonventionen" – "Convention on Children's Rights") are signals for pledges.

7 Related Work

In the field of political sciences, the elections that we consider have been extensively studied by Håkansson and Naurin (2016), Lindvall et al. (2020) and Adhikari et al. (2020). Moreover, applying NLP methods to the analysis of political parties' statements has recently developed into an active field of research, with various groups investing in the creation of dedicated corpora and in their annotation for specific purposes.

A well-known example is the Manifesto Project (MP) (Volkens et al., 2012; Merz et al., 2016), which collects electoral programs from more than 50 countries for democratic elections since 1945, making it a notable initiative within the field. It provides data on different manifesto aspects in several countries and over time. Recently, the Comparative Party Pledges Project (CPPP) of Naurin et al. (2019) has added detailed qualitative coding of what exactly pledges are made of (Naurin and Thomson, 2020).

Subramanian et al. (2018) study the MP data, addressing the identification of fine- vs. coarse-grained positions taken by political parties. Despite the different classification task, similarly to our study, they adopt hierarchical models that encode the texts' structure, finding that considering contextual information improve the models' performance. However, they train bi-LSTM networks from scratch, while we rely on pre-trained BERT language models.

Bilbao-Jayo and Almeida (2018) also work on the MP corpus, applying multi-input Convolutional Neural Networks (CNN) that take into account the statements' context, analogously to our study. They seek to classify the texts according to seven topics corresponding to general areas of interest.

We are partially using the same data as the MP, as we study Swedish manifestos included in that data set. However, we are specifically interested in the identification of election pledges. This is similar to the task studied by Subramanian et al. (2019a). They focus on eleven Australian federal election cycles and distinguish rhetorical (broad) from detailed (narrow) pledges. The annotation of the Swedish texts considers this distinction, while the annotated Indian texts of our corpus do not (Section 2). Subramanian et al. (2019a) use a bidirectional Gated Recurrent Unit (biGRU) to carry out the prediction over ordinal classes.

From a methodological point of view, our approach is related to that of Abercrombie et al. (2019), which also uses BERT. They work on motions tabled in the UK Parliament and find that BERT effectively detects specific categories of proposals in the politicians' speeches.

Concerning the MTL methods, our study is analogous to that of Subramanian et al. (2019b). They consider texts from the 2016 Australian election and propose a new annotation scheme for different *speech acts*. They also perform the classification task using biGRU networks with ELMo embeddings (Peters et al., 2018), relying on a MTL framework in which the auxiliary task is the party prediction: this is also one of our experimental conditions.

8 Conclusion

This paper proposes deep neural models that combine pre-trained language models and trainable attention mechanisms to identify election pledges in party manifestos. We find that these models clearly outperform a non-neural baseline. Even in ZSL conditions (with some contribution by the MTL methods), the performance of the multilingual models indicates that we could even identify pledges in low-resource languages.

Finally, we gained some insight into election pledges' linguistic profile. They are self-contained statements, independent of the context in which they appear. They are likely to be characterized by formulaic expressions that express commitment, intentions, and temporal terms concerning concrete topics. These results result from close interdisciplinary cooperation between two different scientific communities: political scientists and NLP researchers. Pledge identification is the first step for future downstream NLP tasks. For example, the fine-grained study of topics, biases, and the temporal evolution of election pledges in the theoretical framework of political science, which is typically interested in societal developments and explanations such as pledge fulfillment and power distribution in democracies.

9 Ethical Considerations

The data we release are publicly available political manifestos. The texts are not harmful and do not contain personal information. The annotation and the relative classification task do not raise privacy concerns.

References

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

Pankaj Adhikari, Sania Mariam, and Robert Thomson. 2020. Indian parties' election pledges and pledge fulfilment. *APSA Online Annual Conference*.

Aritz Bilbao-Jayo and Aitor Almeida. 2018. Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, 14(11):1550147718811827.

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*.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Nicklas Håkansson and Elin Naurin. 2016. Promising ever more: An empirical account of swedish parties' pledge making during 20 years. *Party Politics*, 22(3):393–404.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. *arXiv preprint arXiv:1911.06194*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Johannes Lindvall, Hanna Bäck, Carl Dahlström, Elin Naurin, and Jan Teorell. 2020. Sweden's parliamentary democracy at 100. *Parliamentary Affairs*, 73(3):477–502.

Nicolas Merz, Sven Regel, and Jirka Lewandowski. 2016. The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):2053168016643346.

Elin Naurin, Terry J Royed, and Robert Thomson. 2019. *Party mandates and democracy: Making, breaking, and keeping election pledges in twelve countries*. New Comparative Politics.

Elin Naurin and Robert Thomson. 2020. The fulfilment of election pledges. In *Research Handbook on Political Representation*. Edward Elgar Publishing.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What's in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. *arXiv preprint arXiv:1805.02823*.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019a. Deep ordinal regression for pledge specificity prediction. *arXiv preprint arXiv:1909.00187*.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019b. Target based speech act classification in political campaign text. *arXiv preprint arXiv:1905.07856*.

Robert Thomson, Terry Royed, Elin Naurin, Joaquín Artés, Rory Costello, Laurenz Ennser-Jedenastik, Mark Ferguson, Petia Kostadinova, Catherine Moury, François Pétry, et al. 2017. The fulfillment of parties' election pledges: A comparative study on the impact of power sharing. *American Journal of Political Science*, 61(3):527–542.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Andrea Volkens, Onawa Lacewell, Pola Lehmann, Sven Regel, Henrike Schultze, and Annika Werner. 2012. The manifesto data collection. manifesto project (mrg/cmp/marpor). *Wissenschaftzentrum Berlin (https://manifestoproject. wzb. eu/)*.