Deep Ordinal Regression for Pledge Specificity Prediction

Shivashankar Subramanian   **Trevor Cohn**   **Timothy Baldwin**
School of Computing and Information Systems
The University of Melbourne
shivashankar@student.unimelb.edu.au   {t.cohn,tbaldwin}@unimelb.edu.au

Abstract

Many pledges are made in the course of an election campaign, forming important corpora for political analysis of campaign strategy and governmental accountability. At present, there are no publicly available annotated datasets of pledges, and most political analyses rely on manual analysis. In this paper we collate a novel dataset of manifestos from eleven Australian federal election cycles, with over 12,000 sentences annotated with specificity (e.g., rhetorical vs. detailed pledge) on a fine-grained scale. We propose deep ordinal regression approaches for specificity prediction, under both supervised and semi-supervised settings, and provide empirical results demonstrating the effectiveness of the proposed techniques over several baseline approaches. We analyze the utility of pledge specificity modeling across a spectrum of policy issues in performing ideology prediction, and further provide qualitative analysis in terms of capturing party-specific issue salience across election cycles.

1 Introduction

Election manifestos play a critical role in structuring political campaigns. Campaign communication can influence a party's reputation, credibility, and competence, which are primary factors in voter decision making (Fernandez-Vazquez, 2014). Among the various campaign-related functions fulfilled by manifestos (Eder et al., 2017), perhaps the most important is the contract they represent between parties and voters in terms of pledges and prioritisation of political issues (Royed et al., 2019). Political scientists have long studied how specific pledges translate into government programs and actual policy (Royed, 1996; Thomson, 2001; Naurin, 2011; Schermann and Ennser-Jedenastik, 2014). Other work relates specific pledges to the issue clarity of a political party

through selective emphasis, which complements salience theory (Robertson et al., 1976; Budge and Farlie, 1983; Praprotnik, 2017). For example:

> *we commit ... 30 per cent tax rebate or cash benefit on the cost of private health insurance premiums*

conveys the party's support for private health insurance, and is more verifiable than:

> *we will improve the health system.*

Issue clarity has also been shown to be influenced by a party's ideological position and its role in government (Praprotnik, 2017).

Although pledge specificity prediction is an important task for the analysis of party position, priorities, and post-election policy framing, to date, almost all research has relied on manual analysis. Subramanian et al. (2019) is a recent exception to this, in performing speech act classification over political campaign text, where the class schema includes the distinction between specific and vague pledges (binary specificity class).

In this paper, we perform fine-grained pledge specificity prediction, which is more expressive than binary levels (Li et al., 2016; Gao et al., 2019). We use a class schema proposed by Pomper and Lederman (1980) as detailed in Table 1, which captures seven levels of specificity, forming a non-linear increasing order of commitment and specificity (Pomper and Lederman, 1980). Given the non-linear nature of the scale, we use deep ordinal regression models for this task, with distributional loss (Imani and White, 2018), where we model the output as a uni-modal distribution (Beckham and Pal, 2017). Our goal is to capture the intuition that a pledge with specificity level k, has higher commitment than all the levels $< k$, producing a smoothly varying prediction over the ordinal classes. This can be modeled as a uni-modal

distribution which has a probability mass that decreases on both sides of the most probable class. Lastly, as it is expensive to obtain large-scale annotations, in addition to developing a novel annotated dataset, we also experiment with a semi-supervised approach by using unlabeled text.

The contributions of this paper are as follows: (1) we develop and release a dataset[1] for fine-grained pledge specificity prediction based on election manifestos covering eleven Australian federal election cycles (1980–2016), from the two major political parties — Labor and Liberal; (2) we propose to use deep ordinal regression models for the prediction task, and evaluate the model under sparse supervision scenarios using the teacher–student framework; and (3) we evaluate the utility of pledge specificity towards ideology prediction, and provide further qualitative analysis by correlating model predictions with party-specific issue salience across major policy areas.

2 Related Work

Political manifesto text analysis is a relatively novel application, at the intersection of Political Science and NLP. Research has focused primarily on fine-grained policy topic classification and overall ideology prediction tasks (Volkens et al., 2017; Verberne et al., 2014; Zirn et al., 2016; Subramanian et al., 2018). Most work dealing with pledge specificity analysis in manifestos has been based on manual analysis, as outlined in Section 1.

Specificity is a pragmatic property of text which has been studied across various fields of research. In cognitive linguistics, Dixon (1987) showed that specificity of information in text impacts reading comprehension speed. In Political Science, it has been used to analyze salience, party position and post-election policy framing (see Section 1). There has also been research on the association between text specificity and communication style. In terms of automated specificity analysis, Cook (2016) found specificity in the context of congressional hearings to vary between speakers belonging to the same vs. different ideologies. Namely, it was shown that specificity increases as the ideological distance between the committee chair and the witness decreases. Subramanian et al. (2019) addressed two levels of pledge specificity, as part of speech act classification task. Specificity has

also been studied in news (Louis and Nenkova, 2011) and classroom discussion domains (Luo and Litman, 2016; Lugini and Litman, 2017).

These studies have dealt with a restrictive coarse-level analysis (2–3 categories), whereas a fine-grained scale better captures and allows for comparison of election manifestos (Pomper and Lederman, 1980). Gao et al. (2019) was the first attempt at fine-grained text specificity prediction, in the context of social media posts. Here, we target the novel task of fine-grained pledge specificity prediction, which can be used in a range of downstream applications, including capturing party priorities (salience) and ideological position across election cycles.

All the text specificity analysis work in NLP has modeled the task as classification or regression. As the 7-step pledge specificity levels used in this research (Pomper and Lederman, 1980) do not form a single real-valued scale, we model it as an ordinal regression task. Some examples of ordinal regression tasks include sentiment rating prediction (Rosenthal et al., 2017), stages of disease prediction (Gentry et al., 2015), and age prediction (Eidinger et al., 2014). Recent work has shown that adding a distributional (auxiliary) loss alongside a regression loss, and using expectation to obtain the predicted value (Imani and White, 2018), provides label smoothing and improves regression performance (Gao et al., 2017). Approaches based on a uni-modal probability distribution (e.g., Poisson) as output (da Costa et al., 2008; Beckham and Pal, 2017) can be seen as related to the former approach (Imani and White, 2018) where the discrete probability mass function replaces the histogram density. We propose to use a uni-modal distributional loss-based ordinal regression for pledge specificity prediction.

Secondly, as it is difficult to obtain large amounts of labeled data, existing approaches have used semi-supervised learning (Li and Nenkova, 2015; Subramanian et al., 2019). Here we use a cross-view training approach (Clark et al., 2018; Subramanian et al., 2019), where we enforce consensus between the intermediate class distributions or the final real-valued output.

3 Pledge Specificity Dataset

We annotated 22 election manifestos from the Australian Labor and Liberal parties, covering eleven Australian federal election cycles from

---

Category	Definition	Example	#
Not a pledge	Provide facts; greetings; approval or criticism of policies	*We have modernised Australia's industrial relations system, particularly through the 1996 Workplace Relations Act*	1
Rhetorical pledge	Based on moral values and applies to all irrespective of the party	*We will put our country first*	2
General pledge	Specify intangible goals, and also not the ways to achieve them	*Labor will build a stronger and more productive economy*	3
Continuity pledge	Commit to the maintenance of currently functioning policy	*We will retain the voluntary health insurance system which now covers the great majority of Australians*	4
Goal pledge	Provide tangible outcomes and goals, without providing the means to achieve them	*A Shorten Labor Government will create 2000 jobs in Adelaide*	5
Action pledge	Provide means to achieve the objective, but don't reveal specific details	*We pledge to support effective voluntary family planning, and to recognize officially the link between social and economic development and the willingness of the individual to limit family size*	6
Detailed pledge	Provide clear details of action to achieve an objective	*A re-elected Coalition Government will invest $1 million to support the Exeter Community Precinct*	7

Table 1: Pledge specificity category, definition, example manifesto sentence, and the corresponding specificity value (#).

Specificity	# Sentences	%	Avg. Length
1	8165	67.00	19.5
2	423	3.47	22.0
3	950	7.80	23.4
4	300	2.46	24.6
5	473	3.88	24.8
6	901	7.39	25.2
7	973	7.99	28.5

Table 2: Distribution and length statistics across specificity categories.

Figure 1: Average pledge specificity of Labor and Liberal parties over different election cycles

1980–2016. The dataset has 12,185 sentences annotated with seven levels of specificity (Pomper and Lederman, 1980). See Table 1 for class definitions and an example of each class. We obtained annotations using the Figure Eight crowd-sourcing platform. For each sentence we provided the previous two sentences from the manifesto (as context), the party which published the manifesto, election year, and incumbent and opposition party details. Each sentence was annotated by at least 3 workers after passing quality control (at least 70% accuracy on test questions). After obtaining annotations, the label which has the highest confidence score is chosen for each sentence. Confidence is the level of agreement between multiple contributors, weighted by the contributors' trust scores. Overall agreement based on the Krippendorf's Alpha is $\alpha = 0.58$, indicating moderate agreement
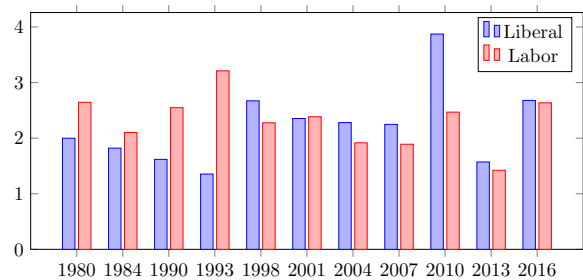
(Artstein and Poesio, 2008; Krippendorff, 2011), on par with related studies (Gao et al., 2019). The class distribution in the final dataset is given in Table 2, alongside the average sentence length in tokens. It can be seen that more specific pledge categories have higher average length. Average specificity values of Labor and Liberal party manifestos across elections are given in Figure 1. The length of the manifesto (in terms of number of sentences) influences average specificity values, with exceptions such as the Liberal party's 2010 election manifesto which is the shortest document but has the highest average pledge specificity value. Further detailed analysis, to decipher the pledge specificity trends in general, is a potential task for future work.

4 Proposed Approach

4.1 Base Model

We first obtain representations for each sentence via a sequence of word embeddings, to which we apply a bidirectional GRU ("biGRU": Cho et al. (2014)), and concatenate the final hidden state of both the forward and backward GRUs, $\mathbf{h}_i = \left[\overrightarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{h}}_i \right]$. Rather than using a linear activation layer for the output, we study the effect of learning a distribution over ordinal classes, and using an expectation layer to get the final prediction, which we now expound upon.

4.2 Distributional Loss

Let us assume that the continuous target variable Y is normally distributed, conditioned on inputs $x \in \mathbb{R}^d$, $Y \sim \mathcal{N}(\mu = f(x), \sigma^2)$ for a fixed variance $\sigma^2 > 0$. In regression, the maximum likelihood function f for n samples $\{x_i, y_i\}$ corresponds to minimizing l_2 loss, such that $f(x) = \mathbb{E}(Y|x)$. Alternatively, we can learn a categorical distribution (q_x) over the ordinal classes $\mathcal{Y}$, and use the expected value as the prediction, $f(x)$ (Rothe et al., 2018). In this work, we follow the latter method, but parameterise the categorical distribution based on uni-modal probability distribution, a technique which has been shown to perform well for ordinal regression tasks (Beckham and Pal, 2017). This modification converts the problem to a more difficult (multi-task) problem, that promotes generalization and reduces over-fitting (Imani and White, 2018). The overall objective is to jointly minimize the squared loss for the regression task ($\mathcal{L}_S$), and cross-entropy for the distributional loss over $\mathcal{Y}$ ($\mathcal{L}_D$), based on the objective $\mathcal{L}_J = \alpha \mathcal{L}_S + \mathcal{L}_D$, where the hyper-parameter α is tuned using a validation set. We experiment with different distributions in generating the intermediate representations q_x, including categorical (as a baseline approach, see Section 5: Beckham and Pal (2016); Gao et al. (2017); Rothe et al. (2018)), discrete uni-modal (Binomial and Poisson: Beckham and Pal (2017)), and truncated Gaussian (Imani and White, 2018). The final prediction is obtained using expectation, which has been shown to be effective for various regression tasks in the vision domain. Here we study the use of uni-modal distributional loss-based ordinal regression approaches (Beckham and Pal, 2017; Imani and White, 2018) for text specificity analysis (Section 5 has re-

sults demonstrating its superiority over the other choices). We detail the different ways to obtain q_x, and the corresponding loss functions $\mathcal{L}_D$ below, and provide an overall summary in Figure 2.

4.2.1 BINOMIAL

With the biGRU model, we estimate the parameter (p) of the Binomial distribution (with a sigmoid output), based on which the distribution over classes can be obtained via the probability mass function,

$$p(k; K-1, p) = \binom{K-1}{k} p^k (1-p)^{K-1-k},$$

$k \in \{0, 1, \ldots K-1\}$, where $p \in [0, 1]$. As the final layers (post sigmoid) are under-parametrized, we have a softmax layer with τ after obtaining the probability masses,

$$q_k = \frac{\exp(\phi_k/\tau)}{\sum_{j=0}^{K-1} \exp(\phi_j/\tau)},$$

where $\tau \sim \text{SoftPlus}(\tau')$, and τ' is learned by the deep net, conditioned on the input (x). We then have an expectation layer to obtain the final output $f(x)$. Output of the softmax layer is fit to the one-hot encoded ordinal classes for each input ($\mathbf{y}$), by minimizing the cross-entropy loss ($\mathcal{L}_{\mathcal{D}\text{BINOMIAL}}$).

4.2.2 POISSON

POISSON is similar to the binomial case, in that we obtain the parameter (λ) of the Poisson distribution using the biGRU, with a SoftPlus activation. We then use the probability mass function of the Poisson distribution to get the probabilities over different classes, $k \in \mathcal{Y}$,

$$p(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

which is again passed through a softmax layer to obtain q_x, fit by minimizing cross-entropy loss ($\mathcal{L}_{\mathcal{D}\text{POISSON}}$), and an expectation layer is used to obtain the final prediction.

4.2.3 Gaussian (GAUSS)

To compute $\mathbb{E}(Y|x)$ (μ of the Gaussian), here we fit the intermediate distribution q_x directly to histogram density of a truncated Gaussian distribution with support $[1, K]$ (target distribution: p_*). We achieve this by learning a prediction distribution with the biGRU model, $q_x : \mathcal{Y} \to [0, 1]$. For this, the ordinal label of training instances is transformed into a truncated Gaussian PDF. The mean
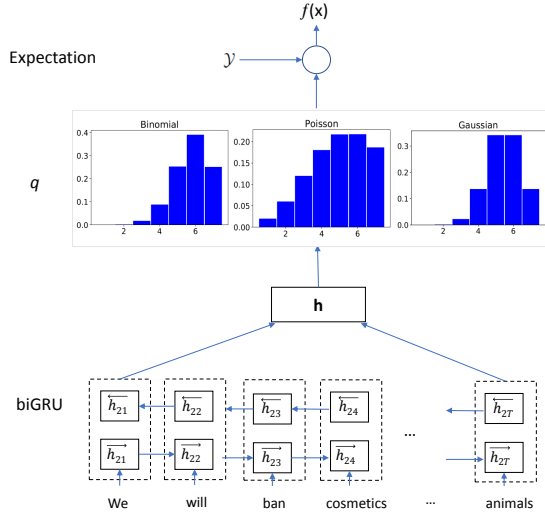
Figure 2: Illustration of the model architecture, comprising a biGRU over sentence tokens, to compute the parameter of one of the three distributions: p for Binomial (with $n = K-1$) and λ for Poisson. *Pmf* of these distributions are then used to define a categorical distribution over the ordinal classes $\mathcal{Y}$. For learning, we use the categorical cross-entropy against the gold *one-hot* y, as well as the squared error $(y - f(x))^2$, where $f(x)$ is given by expectation taken under the predicted categorical distribution q. Gaussian uses a different mechanism, as described in Section 4.2.3 where the categorical distribution (q) is predicted directly using a K-dimensional softmax output, and the cross-entropy is computed between q and a Gaussian histogram density centred at $\mu = y$, discretised by way of integration of the PDF between adjacent label indices.

μ for this Gaussian is the target y of each datapoint, with fixed variance σ^2, which we set to the radius of the bins in $\mathcal{Y}$ (1 in this case). The CDF for the chosen target distribution is computed as $\frac{1}{2}(1 + \text{erf}(\frac{x-\mu}{\sigma\sqrt{2}}))$ and p_* is obtained for each class, $k \in \mathcal{Y}$, as,

$$\frac{1}{2}\left(\text{erf}(\frac{k - \mu}{\sigma\sqrt{2}}) - \text{erf}(\frac{k - 1 - \mu}{\sigma\sqrt{2}})\right).$$

This formulation allows efficient computation of divergence between p_* and q_x for optimization, which results in cross-entropy minimization ($\mathcal{L}_{\text{DGAUSS}}$: Imani and White (2018)). Note that the training target p_* is uni-modal, and no constraints are *explicitly* enforced on the shape of q_x.

4.3 Incorporating Context

We incorporate context in the form of information from adjacent sentences following the approach of Liu et al. (2017): for each training sentence, we

use the predicted (intermediate) probability distribution across ordinal classes of the previous L sentences as context. A new biGRU model is trained with the sentence and the additional contextual information, concatenated to $\mathbf{h}_i$. We refer to this model as biGRU$_{\text{ORD+CONTEXT}}$. In the test phase, biGRU$_{\text{ORD}}$ provides contextual information, and the newly trained model (biGRU$_{\text{ORD+CONTEXT}}$) is used to predict the test sentence output.

4.4 Semi-supervised Learning

As it is expensive to get large-scale specificity annotations we employ a cross-view training approach (Clark et al., 2018; Subramanian et al., 2019) for semi-supervised learning, which can leverage additional unlabeled text. Cross-view training is a kind of teacher–student method, whereby the model "teaches" a "student" model to classify unlabelled data. The student has a restricted view over the data, e.g., through the application of noise (Sajjadi et al., 2016; Wei et al., 2018). We use biGRU$_{\text{ORD+CONTEXT}}$ with word-level dropout and zero vector set to contextual information as the auxiliary model. This procedure regularizes the learning of the teacher to be more robust, as well as increasing its exposure to unlabeled text. We augment our dataset with over 32k sentences from UK and US election manifestos released from the same time period. On these unlabeled examples, the model's output is used to fit the auxiliary model by enforcing consensus in their predictions. This consensus loss $\mathcal{L}_{\text{U}}$ is added to the supervised training objective ($\mathcal{L}_J$). Under the semi-supervised setting, we evaluate the following approaches:

MSE: use the final regression output of the teacher model ($f(x)$) to fit an auxiliary model, thereby enforcing consensus using a squared loss, MSE($\mathbb{E}_{q_\theta}[\mathcal{Y}], \mathbb{E}_{q_\omega}[\mathcal{Y}]$) where $\mathcal{Y}$ is a fixed class vector; denoted as "$\mathcal{L}_{\text{UMSE}}$".

KLD: an intermediate distribution over targets $q_\theta(\mathcal{Y}|s)$ is used to fit an auxiliary model, $q_\omega(\mathcal{Y}|s)$, by minimising the Kullback-Leibler (KL) divergence, KL($q_\theta(\mathcal{Y}|s), q_\omega(\mathcal{Y}|s)$); denoted as "$\mathcal{L}_{\text{UKLD}}$".[2]

EMD: $q_\theta(\mathcal{Y}|s)$ is again used to fit the auxiliary model, $q_\omega(\mathcal{Y}|s)$, by minimising the earth

---

[2]This is the closest setting to Subramanian et al. (2019), which minimizes KL divergence between output distributions in a classification setting. But the overall objective is different in our case, in that we have an expectation layer over q to obtain the target regression output.

Category	Approach	MMAE	ρ
	Majority	3	-
	Length	-	0.21
	Speciteller	-	0.18
	NN_{REG}	2.05	0.33
	$biGRU_{REG}$	1.83	0.47
	$biGRU_{CLASS}$	2.17	0.40
	$biGRU_{REG_{l_1}}$	1.99	0.46
Ordinal	$biGRU_{CATEGORICAL}$	1.80	0.48
Ordinal	$biGRU_{BINOMIAL}$	1.78	0.48
Ordinal	$biGRU_{POISSON}$	1.90	0.41
Ordinal	$biGRU_{GAUSS}$	1.72	**0.49**
Ordinal	$biGRU_{GAUSS + CONTEXT}$	**1.71**	**0.49**

Table 3: Specificity prediction performance; the best approach is given in bold.

mover's distance, $EMD(q_\theta(\mathcal{Y}|s), q_\omega(\mathcal{Y}|s))$; denoted as "$\mathcal{L}_{UEMD}$". EMD is defined as $EMD(q_\theta, q_\omega) = \frac{1}{K}^{\frac{1}{l}} \|cmf(q_\theta) - cmf(q_\omega)\|_l$, where $cmf(\cdot)$ is the cumulative mass function for the predicted (intermediate) probability distribution q, and we use $l = 2$. EMD considers distance between classes, and is more suitable for ordinal tasks (Hou et al., 2016).

5 Experimental Results

To evaluate model performance we use macro-averaged mean absolute error (MMAE: Rosenthal et al. (2017)) given the class imbalance, and Spearman's ρ. MMAE is given as $\frac{1}{|K|} \sum_{j=1}^{|K|} \frac{1}{|S_j|} \sum_{x_i \in S_j} |f(x_i) - y_i|$, where S_j denotes the subset of instances annotated with (true) ordinal class j. We consider the following baselines:

Majority: assign the majority class in the training set to all test instances.

Length: use sentence length as the specificity score.

Speciteller: co-training model of Li and Nenkova (2015), used by Cook (2016) for congressional hearings specificity analysis.

NN_{REG}: bag-of-words term-frequency representation, fed into a feed-forward neural network model (Gao et al., 2019).

$biGRU_{REG}$: biGRU model trained with a mean squared loss objective.

$biGRU_{CLASS}$: biGRU model trained with a cross-entropy objective (Subramanian et al., 2019).

$biGRU_{REG_{l_1}}$: biGRU regression model with mean absolute error objective (l_1 loss).

All the baseline and proposed biGRU models

use ELMo embeddings (Peters et al., 2018). The regression models minimize l_2 loss, unless otherwise specified. We compare the average performance across five runs with an 80:20 train:test split. We randomly choose 10% of instances from the training set as validation data. We compare the baseline approaches with our proposed ordinal approaches, which have an intermediate distributional loss in conjunction with the final prediction loss: $biGRU_{GAUSS}$ (Section 4.2.3), $biGRU_{BINOMIAL}$ (Section 4.2.1), or $biGRU_{POISSON}$ (Section 4.2.2). We also evaluate $biGRU_{CATEGORICAL}$, where the softmax layer is fitted to one-hot encoded class labels (Gao et al., 2017; Rothe et al., 2018). Note that this is not uni-modal.

Gao et al. (2019) used a combination of bag-of-words representation, surface features, social-media-specific features (eg., Tweet mentions), and emotion-related features, with a support vector regression model which minimizes squared loss. Social media and emotion-related attributes are not relevant to our data, and other surface features did not provide improvements. Hence we show the performance of the bag-of-words representation with squared loss objective (NN_{REG} in Table 3). From the results in Table 3, we can see that sequential models with ELMo embeddings (biGRU) perform better than neural bag-of-words models (NN_{REG}). The l_2 regression model ($biGRU_{REG}$) performs better than l_1 regression ($biGRU_{REG_{l_1}}$) and classification ($biGRU_{CLASS}$).

With respect to deep ordinal approaches, $biGRU_{POISSON}$ performs better than classification ($biGRU_{CLASS}$), but does not improve upon regression ($biGRU_{REG}$). The Binomial performs better than the Poisson, consistent with previous work (Beckham and Pal, 2017; da Costa et al., 2008). $biGRU_{CATEGORICAL}$ performs better than $biGRU_{REG}$, but not over unimodal approaches ($biGRU_{BINOMIAL}$, $biGRU_{GAUSS}$). Overall, the model which fits intermediate distribution to a truncated Gaussian (histogram density) target distribution provides the best performance. It gives over 6% improvement in terms of MMAE, and over 4% in ρ, compared to $biGRU_{REG}$. Adding context to $biGRU_{GAUSS}$ ($biGRU_{GAUSS + CONTEXT}$) provides a slight reduction in error.

5.1 Semi-supervised Learning

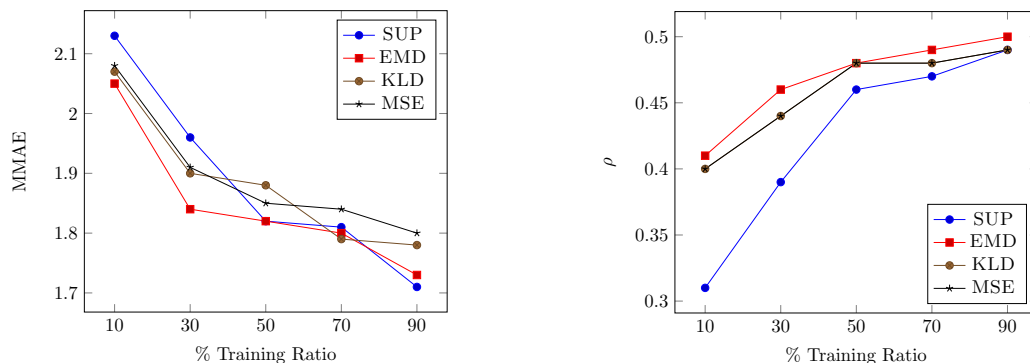We next compare the performance of $biGRU_{GAUSS + CONTEXT}$ (SUP) and the semi-

Figure 3: Prediction performance across different training ratios. Note that 90% = all the training data, as 10% is used for validation. The supervised ordinal model (SUP) and semi-supervised teacher–student models (MSE, KLD, EMD, given in Section 4.4) are compared on MMAE and Spearman's ρ.

supervised extensions of it (Section 4.4) which leverage additional unlabeled data: minimizing $\mathcal{L}_{\text{UMSE}}$ (MSE), $\mathcal{L}_{\text{UKLD}}$ (KLD), and $\mathcal{L}_{\text{UEMD}}$ (EMD). The amount of labeled data in the training split is varied from 10% to 90%. Results are presented in Figure 3 for MMAE and ρ. From the results, semi-supervised approaches provide large gains in terms of both MMAE and ρ, especially when training with fewer instances, ratio $\leq 30\%$. Overall, the semi-supervised learning approach which minimizes EMD performs best across all training ratios compared to both supervised and other semi-supervised approaches. It provides .10 and .06 absolute improvements in ρ under sparse supervision scenarios (10% and 30% of training data, resp.). Even under richer supervision settings ($\geq 70\%$), it provides higher ρ.

6 Political Analysis Using the Models

Political scientists utilize pledge specificity for a variety of applications (see Section 1). Here, we extrinsically evaluate our specificity model using two tasks related to campaign strategy: (1) party position or ideology prediction (Section 6.1), and (2) issue salience analysis (Section 6.2). For both tasks, we compare the use of pledge specificity across policy issues vs. a count-based representation of policy mentions.

6.1 Ideology Prediction

Estimating the manifesto-level ideology score on the left–right spectrum using sentence-level policy topic annotations is a popular task (Slapin and Proksch, 2008; Lowe et al., 2011; Däubler and Benoit, 2017; Subramanian et al., 2018), for which

the policy scheme provided by CMP (Volkens et al., 2017) is commonly used. It has 57 political themes, across 7 major categories. Among those approaches, the RILE index is the most widely adopted (Merz et al., 2016; Jou and Dalton, 2017), and has been shown to correlate highly with other popular scores (Lowe et al., 2011). RILE is defined as the difference between count of (pre-determined) right and left policy theme mentions across sentences in a manifesto (Volkens et al., 2013). Here we evaluate the effectiveness of using the proposed specificity modeling across those policy issues, compared to using RILE-based party position scores (Volkens et al., 2013).

We compute the specificity weight (Pomper and Lederman, 1980) from the average specificity score across sentences, $\frac{1}{|I|} \sum_{S_i \in I} \text{Spec}(S_i)$ for each policy issue (I). With specificity weight as the basic feature, we also model global signals such as party coalition and temporal dependencies across elections, which can enforce smoothness in manifesto positions (Greene, 2016; Subramanian et al., 2018) based on probabilistic soft logic.

6.1.1 Probabilistic Soft Logic

To address this, we propose an approach using hinge-loss Markov random fields ("HL-MRFs"), a scalable class of continuous, conditional graphical models (Bach et al., 2013). These models can be specified using Probabilistic Soft Logic ("PSL": Bach et al. (2017)), a weighted first order logical template language. An example of a PSL rule is $\lambda : \texttt{P(a)} \land \texttt{Q(a,b)} \rightarrow \texttt{R(b)}$, where $\texttt{P}$, $\texttt{Q}$, and $\texttt{R}$ are predicates, $\texttt{a}$ and $\texttt{b}$ are variables, and λ is the weight associated with the rule. PSL uses soft

truth values for predicates in the interval $[0, 1]$. The degree of ground rule satisfaction is determined using the Lukasiewicz t-norm and its corresponding co-norm as the relaxation of the logical AND and OR, respectively. The weight of the rule indicates its importance in the HL-MRF probabilistic model, which defines a probability density function of the form:

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\left(-\sum_{r=1}^{M} \lambda_r \phi_r(\mathbf{Y}, \mathbf{X})\right),$$

where $\phi_r(\mathbf{Y}, \mathbf{X}) = \max\{l_r(\mathbf{Y}, \mathbf{X}), 0\}^{\rho_r}$ is a hinge-loss potential corresponding to an instantiation of a rule, and is specified by a linear function l_r and optional exponent $\rho_r \in \{1, 2\}$.

6.1.2 PSL Model

Here we elaborate on our PSL model based on manifesto content-based features (specificity weight across 57 policy issues), coalition information, and temporal dependencies. Our target `pos` (left–right position) is a continuous variable $[0, 1]$, where 1 indicates an extreme right position, 0 denotes an extreme left position, and 0.5 indicates center. We also model the social and economic positions explicitly (`socpos` and `econpos`), which influence the overall `pos`. Each instance of a manifesto, its party affiliation and policy issues, are denoted by the predicates `Manifesto`, `Party` and `Policy`. Other predicates are given as follows:

Specificity weight of each policy issue in the given manifesto (`Specw`).

Relative specificity scale: ratio of specificity weight for each policy issue given a party's manifesto, to maximum specificity weight for the same policy issue across parties from the same country and election (`SpecScale`).

Policy issue mapping: 26 out of the 57 policy themes are categorized as social and economic left–right issues by Benoit and Laver (2007) (`IdeologyMap`).

Coalition: captures the strength of ties between two parties, given by a logistic transformation of the number of times two parties have been in a coalition in the past (`Coalition`).

Temporal dependency between a party's current manifesto position and its previous manifesto position (`PreviousManifesto`).

Representative rules of our PSL model, based on the predicates presented above, are given in Table 4. They include:

Specificity: if a manifesto contains more specific pledges related to social (or economic) left/right policies, then it will more likely be a social (or economic) left/right-aligned manifesto.

Overall position: social and economic position influences the overall position, and this allows the model to place different weights on the influence of social and economic policies on the overall position, which is found to be necessary by Benoit and Laver (2007).

Global signals: coalition and temporal dependencies to enforce smoothness in manifesto positions.

Relative specificity: `SpecScale` of a left (or right) policy during an election amplifies its overall position scores.

6.1.3 Evaluation

We use manifestos from Australia and UK for our analysis. We use data from Voter Survey (Cameron and McAllister, 2019) for Australia and CHES Expert Survey (Bakker et al., 2015) for the UK as the gold-standard party position. A primary step (related to the model given in Section 6.1.2) is to obtain policy topic classification for sentences in each manifesto. If annotations are not available from Volkens et al. (2017), one out of 57 political themes are predicted using the method of Subramanian et al. (2018). Specificity scores of sentences are obtained using the proposed ordinal regression approach (biGRU$_{\text{GAUSS+CONTEXT}}$). Using social, economic and a combined list of left–right policy themes (`IdeologyMap`), and with the RILE formulation, we bootstrap `socpos`, `econpos` and `pos`. We then use the PSL model (Table 4) to recalibrate the scores based on specificity scores and the global signals.

We compare the performance of bootstrapped `pos` (RILE or policy count-based) with the PSL model. Principal component analysis ("PCA": Gabel and Huber (2000)) on the frequency distribution, and projection on its principal component, is used as an additional baseline. Spearman's correlation (ρ) against the gold-standard positions is given in Table 5. Overall, pledge specificity, especially on a relative scale (which differentiates emphasis between parties) provides large gains, and global signals give only mild improvements.

Specificity Weight — Model I
Manifesto(x) ∧ Policy(I) ∧ Specw(x, I) ∧ IdeologyMap(I, social left/right) → socpos(x)
Manifesto(x) ∧ Policy(I) ∧ Specw(x, I) ∧ IdeologyMap(I, economic left/right) → econpos(x)
Overall position — Model II
Manifesto(x) ∧ socpos(x) → pos(x)
Manifesto(x) ∧ econpos(x) → pos(x)
Global signals — Model III
Manifesto(x) ∧ Party(x, a) ∧ Manifesto(y) ∧ Party(y, b) ∧ Coalition(a, b) ∧ pos(x) → pos(y)
Manifesto(x) ∧ Party(x, a) ∧ PreviousManifesto(x, t) ∧ Party(t, a) ∧ pos(t) → pos(x)
Relative specificity — Model IV
Manifesto(x) ∧ Policy(I) ∧ SpecScale(x, I) ∧ IdeologyMap(I, social left/right) → socpos(x)
Manifesto(x) ∧ Policy(I) ∧ SpecScale(x, I) ∧ IdeologyMap(I, economic left/right) → econpos(x)

Table 4: PSL Model: Representative rules. `left/right` in the `IdeologyMap` predicate indicates policy issues mapped to left/right categories, which is implemented as two separate rules — one for left and another for right.

Model	Aus	UK
bootstrapped pos	0.66	0.39
PCA	0.11	0.39
Model I + II	0.63	0.33
Model I + II + III	0.65	0.33
Model I + II + III + IV	**0.71**	**0.45**

Table 5: Spearman's ρ for prediction of party position based on the different models.

Policy Area	**C**	**S**
Health	824.92	905.19
Education	781.15	905.54
Environment	841.89	897.53
Tax	735.97	824.69
Economy	258.30	276.71

Table 6: Log-likelihood with pledges specificity weight (**S**) and count of sentences (**C**) across 57 policy themes as independent variables. Log-likelihood values using **S** are better than **C** across all the policy areas.

6.2 Capturing Issue Salience

For the Australian manifestos (from the Greens, Labor, Liberal, and National parties) we perform a qualitative study of specificity weight across policy themes, by correlating it against the salience of major policy areas given by the Voter Survey (Cameron and McAllister, 2019). Again we compare its utility over the use of counts across policy themes in a manifesto. Using sentences classified with policy themes and specificity scores using our proposed approach, we construct the following |Manifestos| × 57 features — frequency distribution (**C**) and pledge specificity weight (**S**) across policy themes. The features are used as independent variables, and voter survey salience scores across major policy areas — health, education, environment, tax, and economy — are treated as dependent variables. Note that the voter survey scores are available for each party and election cycle across policy areas. We build separate multivariate linear regression models and compare them based on the goodness of fit (log-likelihood). Log-likelihood values are given in Table 6: across all policy areas, pledge specificity better captures salience than a count-based representation.

7 Conclusion and Future Work

In this work we present a new dataset of election campaign texts, annotated with pledge specificity on a fine-grained scale. We study the use of deep ordinal regression approaches using an auxiliary uni-modal distributional loss for this task. The proposed approaches provide large gains in performance under both supervised and semi-supervised settings. Specificity weight across policy issues benefits ideology prediction and also better captures issue salience, compared to the traditional policy theme count-based representation. This aligns with previous studies done based on manual annotations (Praprotnik, 2017). In future work, we aim to expand this study to multiple languages.

Acknowledgements

1737

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18(109):1–67.

Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. 2013. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 32–41.

Ryan Bakker, Catherine De Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2015. Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010. *Party Politics*, 21(1):143–152.

Christopher Beckham and Christopher Pal. 2016. A simple squared-error reformulation for ordinal classification. *arXiv preprint arXiv:1612.00775*.

Christopher Beckham and Christopher Pal. 2017. Unimodal probability distributions for deep ordinal classification. In *International Conference on Machine Learning*, pages 411–419.

Kenneth Benoit and Michael Laver. 2007. Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies*, 26(1):90–107.

Ian Budge and Dennis Farlie. 1983. Party competition: selective emphasis or direct confrontation?: an alternative view with data. *Western European party systems : continuity & change*, pages 267–305.

Sarah Cameron and Ian McAllister. 2019. Australian election study, 1987-2016 trends.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925.

Ian Cook. 2016. *Content and Context: Three Essays on Information in Politics*. Ph.D. thesis, University of Pittsburgh.

Joaquim F Pinto da Costa, Hugo Alonso, and Jaime S Cardoso. 2008. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91.

Thomas Däubler and Kenneth Benoit. 2017. Estimating better left-right positions through statistical scaling of manual content analysis. *Retrieved from* `http://kenbenoit.net/pdfs/text_in_context_2017.pdf`.

Peter Dixon. 1987. The processing of organizational and component step information in written directions. *Journal of memory and language*, 26(1):24–35.

Nikolaus Eder, Marcelo Jenny, and Wolfgang C Müller. 2017. Manifesto functions: How party candidates view and use their party's central policy document. *Electoral Studies*, 45:75–87.

Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179.

Pablo Fernandez-Vazquez. 2014. And yet it moves: The effect of election platforms on party policy images. *Comparative Political Studies*, 47(14):1919–1944.

Matthew J Gabel and John D Huber. 2000. Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science*, pages 94–103.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838.

Yifan Gao, Yang Zhong, Daniel Preotiuc-Pietro, and Junyi Jessy Li. 2019. Predicting and analyzing language specificity in social media posts. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.

AE Gentry, CK Jackson-Cook, DE Lyon, and KJ Archer. 2015. Penalized ordinal regression methods for predicting stage of cancer in high-dimensional covariate spaces. *Cancer informatics*, 14(Suppl 2):201–208.

Zachary Greene. 2016. Competing on the issues: How experience in government and economic conditions influence the scope of parties' policy messages. *Party Politics*, 22(6):809–822.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*.

Ehsan Imani and Martha White. 2018. Improving regression performance with distributional losses. In *Proceedings of the International Conference on Machine Learning*, pages 2162–2171.

Willy Jou and Russell J. Dalton. 2017. Left-right orientations and voting behavior. *Oxford Research Encyclopedia of Politics*.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Technical Report, University of Pennsylvania*.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2281–2287. AAAI Press.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3921–3927.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 605–613.

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

Luca Lugini and Diane Litman. 2017. Predicting specificity in classroom discussion. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–61.

Wencan Luo and Diane Litman. 2016. Determining the quality of a student reflective response. In *The Twenty-Ninth International FLAIRS Conference*.

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).

E. Naurin. 2011. *Election Promises, Party Behaviour and Voter Perception*. Palgrave Macmillan.

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

Gerald M. Pomper and Susan S. Lederman. 1980. *Elections in America: Control and Influence in Democratic Politics*. Longman.

Katrin Praprotnik. 2017. Issue clarity in electoral competition. Insights from Austria. *Electoral Studies*, 48:121–130.

David Bruce Robertson et al. 1976. *A theory of party competition*. Wiley London.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157.

Terry J Royed. 1996. Testing the mandate model in Britain and the United States: Evidence from the Reagan and Thatcher eras. *British Journal of Political Science*, 26(1):45–80.

Terry J Royed, Elin Naurin, and Robert Thomson. 2019. *Making and Keeping Pledges*. New Comparative Politics.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1163–1171.

Katrin Schermann and Laurenz Ennser-Jedenastik. 2014. Coalition policy-making under constraints: Examining the role of preferences and institutions. *West European Politics*, 37(3):564–583.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

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

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Target based speech act classification in political campaign text. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 273–282.

Robert Thomson. 2001. The programme to policy linkage: The fulfilment of election pledges on socio–economic policy in the Netherlands, 1986–1998. *European Journal of Political Research*, 40(2):171–197.

Suzan Verberne, Eva D'hondt, Antal van den Bosch, and Maarten Marx. 2014. Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Andrea Volkens, Judith Bara, Budge Ian, and Simon Franzmann. 2013. Understanding and validating the left-right scale (RILE). In *Mapping Policy Preferences From Texts: Statistical Solutions for Manifesto Analysts*, chapter 6. Oxford University Press.

Andrea Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, and Bernhard Weßels. 2017. *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2017b.* Wissenschaftszentrum Berlin für Sozialforschung.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. 2018. Improving the improved training of Wasserstein GANs: A consistency term and its dual effect. In *Proceedings of the International Conference on Learning Representations*.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, pages 88–93.