Author/s:
Subramanian, Shivashankar

Title:
Natural Language Processing for Improving Transparency in Representative Democracy

Date:
2020

Persistent Link:
https://hdl.handle.net/11343/258659

Natural Language Processing for Improving Transparency in Representative Democracy

by

Shivashankar Subramanian

A thesis submitted in total fulfillment for the
degree of Doctor of Philosophy

in the
School of Computing and Information Systems
Melbourne School of Engineering
THE UNIVERSITY OF MELBOURNE

January 5, 2021

Abstract

Language is the natural medium in politics, hence among several types of data, text is the central artifact to capture political behaviour. In this thesis we focus on automating several political analyses using natural language processing techniques, which can improve the transparency of policy-making and thereby the voters' trust in **representative democracy**. Political scientists have observed that the voters' trust in government is necessary for successful implementation of policies, and in-turn their trust is based on effective implementation of policies and services. In-order to improve the trust, the policy-making process should be more transparent and receptive. The policy-making process typically consists of several stages, and we focus on the two primary stages involving *political parties* and *voters* — policy proposal and its implementation audit.

Specifically, we target three major aspects of policy-making process: (a) analyzing policy proposal during election campaign — what policy goals are spoken about, and specifically in which context, and what promises are made. (b) Post-election policy implementation audit — given the pre-election promises, which sets the expectation of voters, does the government make progress towards those promises. (c) Public advocacy for policy changes — what changes do the voters want. This can be seen as both evaluation of existing policies as well as suggestions for changes. More importantly, active participation of voters in the process reflects their level of trust in the system. We define the individual research targets based on political science literature and automate those using deep learning approaches. We use canonical sources of text for each of the tasks, for example, election manifestos released by political parties (more sources are discussed in Chapter 2).

The challenges involved in this work are multi-fold, starting from defining the task, to dataset creation, to developing suitable models. We hope that this PhD thesis, dealing with political text analysis, will shed light on the available data sources, flavor of tasks at the intersection of both natural language processing and political science, and also the techniques to handle the challenges.

Declaration of Authorship

This is to certify that

- this PhD thesis was written by myself and comprises my original work towards the PhD degree, except where specified in the Preface;
- the work has not been submitted for any professional qualification,
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Shivashankar Subramanian

Date: January 5, 2021

Preface

This thesis is a combination of the following papers:

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin and Julian Brooke (2017). Joint Sentence-Document Model for Manifesto Text Analysis, *In Proceedings of the 2017 Annual Workshop of the Australasian Language Technology Association* (ALTA), pp. 25-33. (Won the Best Paper Award)

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* (NAACL-HLT), pp. 1964—1974.

Shivashankar Subramanian, Timothy Baldwin and Trevor Cohn (2018). Content-based Popularity Prediction of Online Petitions Using a Deep Regression Model, *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 182—188.

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), pp. 273—282.

Shivashankar Subramanian, Trevor Cohn and Timothy Baldwin (2019). Deep Ordinal Regression for Pledge Specificity Prediction, *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), pp. 1729—1740.

Shivashankar Subramanian, Trevor Cohn and Timothy Baldwin (2020). Towards Automatic Tracking of Election Promises. *Under submission to the Computational Linguistics Journal* (CL).

Note that I am a primary author of all the above-mentioned publications. In addition, I have contributed to the following work that are **not** included in my thesis:

Shivashankar Subramanian, Ioana Baldini, Sushma Ravichandran, Dmitriy A Katz-Rogozhnikov, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Kush R Varshney, Annmarie Wang, Pradeep Mangalath and Laura B. Kleiman (2020). A Natural Language Processing System for Extracting Evidence of Drug Repurposing from Scientific Publications. *In Proceedings of the Thirty-Second Annual Conference on Innovative Applications of Artificial Intelligence* (IAAI). (This work was done during my internship at IBM TJ Watson Research).

Shivashankar Subramanian, Timothy Baldwin, Julian Brooke and Trevor Cohn (2017). Pairwise webpage coreference classification using distant supervision. *In Proceedings of the 26th International Conference on World Wide Web Companion* (WWW), pp. 841-842.

Shivashankar Subramanian, Daniel King, Sergey Feldman and Doug Downey (2021). S2AND: A Benchmark and Evaluation System for Author Name Disambiguation. *Under submission to the Joint Conference on Digital Library* (JCDL). (This work was done during my internship at Allen AI).

Acknowledgements

I would like to thank my supervisors, Trevor Cohn and Tim Baldwin, for giving me the opportunity to join their group and for nurturing my research skills. Their support was crucial in different stages of each of my research work — idea conception, asking right research questions, developing the ideas and finally writing the work. It is evident from all my paper reviews, that had mentions of how well the paper was written. They tirelessly motivated me, so I did not settle down for easy options along the way, not just in research but also towards the overall career goals. Their support was instrumental for my timely thesis submission amidst COVID-19. I would also like to thank Julian Brooke (currently at The University of British Columbia), who was my co-supervisor during the first year of PhD, for many thoughtful discussions and feedback related to manifesto analysis research. Many parts of the thesis work has benefited through rich conversations with political science researchers in Melbourne: Jenny Lewis (University of Melbourne), Aaron Martin (University of Melbourne) and Robert Thomson (Monash University).

During my PhD, I participated in two fruitful and productive internships. My first internship was at IBM TJ Watson Research, New York. I would like to thank Kush R. Varshney for giving me the opportunity to intern in his group. That internship provided me an exposure to the field of *AI for Social Good* and the challenges involved in working on social good applications. My second internship was at Allen Institute for AI, Seattle. For this, I would thank Doug Downey, Sergey Feldman, Daniel King and Dan Weld for giving me the opportunity to intern at the Semantic Scholar group and for the support throughout the project. Special thanks to AI2 for making the internship possible in spite of COVID-19. This work gave me an experience of one of the best NLP research industries, and also to build useful research applications with good coding standards.

I would also like to thank the Australian government for supporting my PhD through the Melbourne Research Scholarship. Also, for creating a warm and welcoming environment in the country. I thank the University of Melbourne, Google Australia and my supervisors for conference travel support throughout my PhD. I also had the opportunity to be a part of the NLP group at the University of Melbourne, which provided the required motivation at right times and more importantly helped to ease the stress during the long PhD journey. Last but not least, I'd like to thank my family back in India and my wife who was with me in Australia during my PhD. Their support was instrumental for achieving what I did during the PhD.

Contents

Abstract i

Declaration of Authorship ii

Acknowledgements v

List of Figures viii

List of Tables ix

1 Introduction 1
- 1.1 Motivation . . . 4
- 1.2 Scope . . . 5
- 1.3 Contributions . . . 9
- 1.4 Structure . . . 11

2 Background 12
- 2.1 Policy process . . . 13
- 2.2 Artefacts . . . 17
  - 2.2.1 Election campaign text . . . 17
    - 2.2.1.1 Manifestos . . . 18
    - 2.2.1.2 Other campaign sources . . . 25
  - 2.2.2 Policy-making . . . 28
    - 2.2.2.1 Policy agendas . . . 33
    - 2.2.2.2 Program to policy alignment . . . 35
  - 2.2.3 Online petitions . . . 36
  - 2.2.4 News data . . . 40
  - 2.2.5 Social media discussions . . . 44
- 2.3 NLP challenges . . . 46
- 2.4 Summary . . . 50

3 Research Contribution 51
- 3.1 Manifesto analysis [ALTA 2017, NAACL-HLT 2018] . . . 55
- 3.2 Campaign text speech act analysis [*SEM 2019] . . . 77
- 3.3 Election promises specificity prediction [EMNLP-IJCNLP 2019] . . . 89
- 3.4 Election promises progress tracker [CL 2020] . . . 103

3.5 Online petitions popularity prediction [ACL 2018] 115

4 Conclusion and Future Work **124**
4.1 Research questions 124
4.1.1 Approaches 128
4.1.2 Possible immediate extensions 130
4.2 End-to-end analysis 131
4.3 Reflections and future work 136
4.4 Summary 140

Bibliography **142**

List of Figures

1.1 Interaction between political parties and voters through policy proposal and implementation. Voters respond to policies through their votes and other media such as petitions. . . . 2
1.2 End-to-end overview of the research contributions. . . . 5
1.3 Guardian article on the budget decision related to ABC cuts, which indicates the 2013 election promise made by the Liberal Party of Australia has been broken. 8

2.1 Manifesto snippet taken from the Comparative Manifesto Project for the Liberal Party of Australia, 2001. Manifestos from older elections are made available by scanning the corresponding printed copies, as shown here. Digitized text and its annotations are provided by the project — $\int$ denotes a coherent segment. The starting digit of the fine-grained policy theme denotes its major policy category's code (Table 2.1). See text for code description. . . . 21
2.2 Snippet from Politico story on US Congress passing a bill to carry concealed gun. 28
2.3 Example petition from UK Government Petitions website. . . . 36
2.4 Snippet from UK Government's response to the petition given in Figure 2.3. . . . 37
2.5 One of the Trump's claims verified by PolitiFact. . . . 40
2.6 Sample set of Trump's election promises tracked by PolitiFact. . . . 42
2.7 Snapshot taken from procon.org. Issue discussed is *Should the Death Penalty Be Allowed?*. Pro discussions are given on the left-side, and Con are given on the right-side. . . . 43
2.8 Twitter campaign for a petition written to UK Government. . . . 43

4.1 An example from 2016 Australian election — interaction between Labor and Liberal parties through press-releases. It shows the target party in color, and criticism of other party's policy (interaction) in italics, and the promises made by parties towards broadband communication. . . . 132
4.2 Snapshot taken from Vote Compass: Analysis of Australian political parties' (LIB=Liberals, NAT=Nationals, ALP=Labor Party, GRN=Greens, ON=One Nation) position . . . 133
4.3 Snapshot taken from `http://climatecouncil.org.au` which analyses Australian Labor Party's climate policy. . . . 134
4.4 An example for voters' advocacy in the UK politics wrt Brexit policy. It shows four events (top to bottom) — Theresa May invoking Article 50; 2017 elections where Conservatives promised Brexit, won and formed the government; voters' petition to revoke Article 50 fetched lots of signatures; Government responds that Article 50 cannot be revoked as it breaks their election promise. All the text descriptions are taken from BBC. . . . 135

List of Tables

2.1 Comparative Manifesto Project — major policy themes 19
2.2 Statistics of CMP dataset, 2019 version, showing the *start year* (or election) from which manifestos are available for each country (all continuing to the present day), number of manifestos, coded sentences and parties. 20
2.3 26 out of 57 policy themes coded as *left* and *right* by political scientists. The starting digit of the fine-grained policy theme denotes its major policy's category code (see Table 2.1). 22
2.4 Snippet from 2019 election campaign — The Liberal Party of Australia's media-release 25
2.5 Snippet from 2016 US President debate at the University of Nevada, Las Vegas 26
2.6 Comparative Agendas Project — Major topics 34

3.1 Research contributions summary 52

Chapter 1

Introduction

Language is the primary medium of politics, and is used by various stakeholders in democratic systems. The role of language starts from election campaigns where candidates and political parties articulate their policy positions. Speeches of other constitutional representatives, such as the Queen (or King) in the UK, provide the government with an opportunity to highlight its priorities for the future period, at the first sitting of parliament on forming government. Elected representatives debate legislation and make decisions on behalf of voters, as well as raise questions on policy implementation in order to audit their progress. The public discuss and express their views through social media, and also launch campaigns for the change they wish to see through official channels such as online petitions. Countries debate their stand on key issues in the United Nations, as a forum where leaders discuss major issues in world politics [Jankin Mikhaylov et al. (2017)]. News media report the day-to-day affairs of regional and international political issues, in addition to covering information from many other sources. Analysis of these diverse sources of text, and many others, can provide an understanding of what political actors are saying and writing.

Social scientists have long recognized the value of language for political analysis, and initially relied on manual analysis. The primary problem with that approach is in scaling the analysis to large volumes of political text, where it is difficult to annotate texts in even moderately sized corpora. Moreover it is very expensive to employ annotators for the job, and analyzing large amounts of text manually has been impossible in most cases [Grimmer and Stewart (2013)]. Hence there is a natural need for automated text analysis, as is evident from the proliferation of the *'text-as-data'* field of study, which deals with a broad set of techniques and approaches for

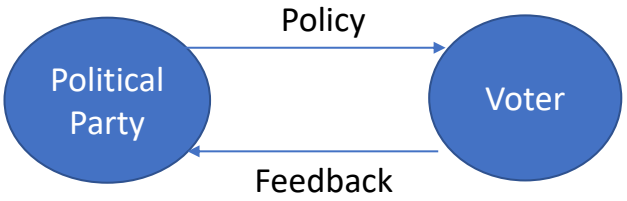


FIGURE 1.1: Interaction between political parties and voters through policy proposal and implementation. Voters respond to policies through their votes and other media such as petitions.

analyzing text generated by various political actors, to answer social science questions. Though there has been a lot of progress in this area, analysis of canonical text sources still largely relies on manual annotation efforts. In this thesis, we develop methods to automate various political science tasks, using natural language processing and machine learning techniques.

Among various activities, policy-making is a pivotal part of governance, and involves interactions between different political actors (more details about the policy-making process and different artefacts generated are discussed in Chapter 2). A highly simplified view of this, showing the functioning of the two main stakeholders of a representative democracy — political parties and voters — is given in Figure 1.1. The text sources generated by this process are both instruments to measure *political accountability* and *voter advocacy*, to which political scientists pay much attention.

An important challenge towards analyzing *political accountability* is to first characterise political parties, given the large amounts of text generated by them. Some of the main characteristics include ideological position, clarity, salience, and dissent within a party, towards different policy issues. An example of this is the **position** of a party in terms of its ideological stance on economic issues, **clarity** of a party's position on economic issues, relative **salience** of economic issues in the party's public stance, and degree of **dissent** on integration policy for immigrants and asylum seekers. These indicators are part of the Chapel Hill expert survey for European political parties [Bakker et al. (2015)]. To perform such analysis, several tiers of text processing are required, including classifying the policy of text, for example, *tax relief for 3.4 million small businesses employing over 7 million Australians* discusses economic policy. The party position analysis also requires stance classification, which in the given example, is to categorize that the party takes a favorable position towards the economy. Furthermore, in order to understand the party's position we should differentiate from other types of statements, such as *Australia will pass this COVID-19 test and emerge stronger on the other side*, which is a belief message.

Distinguishing clarity of goal statements can aid in the analysis of issue clarity of parties, for example, *we will support businesses* versus *we will lower the tax rate to 27.5 per cent this year for businesses*. Here the former statement has a clearer goal compared to the latter one. Salience is about the relative emphasis of issues, while dissent measures differences in stance between individuals in a party. Secondly, performing such analysis over time can help to study the consistency of party position, and can provide the basis for comparison across countries (or states within a country). This further motivates the need to distinguish between past actions and future goals, e.g., *last year, for the first time in 11 years, the budget was returned to balance* vs. *tax relief for over 10 million workers, starting next year*. These analyses help to understand the party's views, position and commitment, which set the expectation for voters in a democracy.

Apart from analyzing individual party characteristics, political scientists study interaction between parties, which can provide insights into whether they put national interests above individual goals, under the broader theme of *constructive democracy*. Further, political accountability can more accurately be measured based on representatives' actions. This includes studying whether the party in power fulfills its promises, whether opposition parties support the policy-making in a way that is consistent with their stance and promises, and what is the influence of coalition parties on policy decisions [Thomson et al. (2017)]. With large amounts of text, especially with increasingly available digitized content on the internet, suitable techniques based on natural language processing and machine learning are required to automate these tasks.

Other than the political parties, voters play a significant role in achieving a healthy democracy. Political scientists argue that the more involved democratic participation of voters, the greater the likelihood of superior social outcomes. Voter participation can help in aggregating policy-level preferences, which is hard to obtain otherwise [Pateman (1970)], and can potentially aid the process of achieving more representative policies. Voters play their role during elections by voting for their candidate or party of choice, based on the party's past performance, their position and promises on various issues, the leader of the party, and so on. Apart from this periodic involvement in the democratic system, their political participation can take many other forms, such as voting on policies in referenda, proposing policy changes, and gathering others' support through suitable campaign strategies — this can be achieved by creating or signing petitions, and participating through contestation (e.g., debates or protests), which are both seen seen as "intense" forms of participation [Weitz-Shapiro and Winters (2008)]. Among these, online petitions can be seen as a constructive way to directly influence the policy agenda, and gain support from thousands of voters quickly. Given the potential effect on democracy, it is

valuable to understand voter stance on issues and their advocacy for policy changes, and design systems to improve their participation in the process [Hale et al. (2018)].

Automating political analysis using text from different sources to answer political science questions has several challenges, starting from defining the task, to dataset creation, to developing suitable models. Since getting access to large amounts of labeled data to train data-hungry models is difficult in any domain, especially in the case of specialized ones like politics, it is necessary to develop approaches which can learn even with small amounts of labeled text. Semi-supervised approaches are a cost-effective way to deal with the challenge [Cardie and Wilkerson (2008)], by leveraging large amounts of unlabeled text, which is available in many scenarios.

1.1 Motivation

Across the globe, in all countries irrespective of their development status, trust in government is decreasing. This applies equally to Australia, as indicated by the latest Australian (2019 Federal) election study conducted by The Australian National University,[1] which found that Australians' satisfaction with democracy has reached its lowest level. They also found that just one-in-four Australians have confidence in their political leaders and institutions. Voters' trust is usually reflected in the level of their participation — voting on the election day, and involvement through other mechanisms such as petitions to let their voices be heard. It should be noted that some countries including Australia have compulsory voting, where voter turnout based analysis may not be appropriate. For people who do **not** vote, a Pew Research study of the 2016 US elections[2] shows the most common reasons were either that *people did not like the election issues* or *did not believe their vote would make a difference*. It should be noted that many voters are unaware of policies proposed by political parties during an election, and the differences between those policies [Kramer (2018)]. Thus, it is necessary to improve voter awareness of party characteristics and policy delivery, including comparative analysis of policies across parties, promises made, and their fulfilment by different parties. Manually analyzing vast amounts of data generated during and after elections is labour intensive, and it is necessary to automate the analysis wherever possible.

Politically-aware voter participation is necessary for a healthy democracy, and online petitions provide an unique opportunity to author policy change proposals, and gather support from other

[1] https://www.anu.edu.au/news/all-news/trust-in-government-hits-all-time-low
[2] https://www.pewresearch.org/fact-tank/2017/06/01/dislike-of-candidates-or-campaign-issu

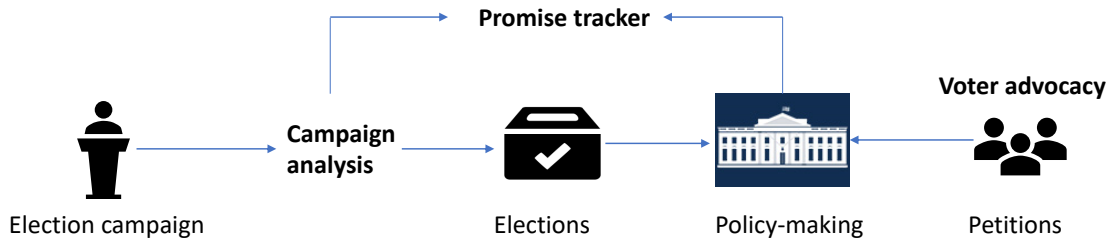


FIGURE 1.2: End-to-end overview of the research contributions.

voters. Other than voting on election day or for policy change referenda, online petitions are a form of voter advocacy for policy changes, and voter participation indicates their trust in the system. Though online petitions do not have a direct effect on policy changes, they can help to increase awareness, participation and attention towards different issues.[3] The majority of petitions receive very little attention in their lifetime, however petition platform design options such as showing number of signatures obtained by a petition [Margetts et al. (2015)], and petition authors' campaign strategies such as reaching out through social media can potentially increase its reach [Yasseri et al. (2017)]. Estimating the reach of a petition at the time of its submission and choosing suitable intervention mechanisms can potentially improve the voters' political participation and thereby create a healthy democracy.

1.2 Scope

Broadly, our focus spans three major topics, which connect political parties and voters (as shown in Figure 1.2): (a) analyzing policy proposals during election campaigns — what policy goals are spoken about, and specifically in which context? (b) Post-election policy implementation audit — given the pre-election promises, which set the expectations of voters, does the government make progress towards those promises? (c) Voter advocacy for policy changes — what changes does the public want? In this section we define our scope of work, which is restricted to the following set of challenges.

CAMPAIGN TEXT ANALYSIS: Election campaigns contain party views, positions and goals, which set the expectations of voters. Analyzing party positions across policy issues, and

[3] https://www.abc.net.au/news/2016-12-15/are-online-petitions-ever-effective/8124388

their campaign-related rhetorical functions can help to improve awareness of voters in election issues.

ELECTION PROMISE TRACKER: The core aspect of political accountability is whether election promises are kept by a party. Promise tracking can provide a consolidated assessment of the government's performance, thereby facilitating voters to keep track of policy-related progress, which can help them to take informed decisions in following elections, or intervene through the available mechanisms for advocacy.

VOTER ADVOCACY: Online petitions provide a unique channel for voter advocacy. It is known that the majority of petitions do not succeed, in terms of reaching the target number of signatures to obtain a response from government, and have their fate decided during the initial few days after they are submitted [Yasseri et al. (2017)]. Hence, predicting popularity at the time of submission provides an opportunity to adapt suitable campaign mechanisms to improve reach among supporters.

We further expound upon the scope as a set of research questions (RQs), and detail how we propose to achieve it in this thesis.

RQ1: How can election campaign text be used to help voters keep track of party views and policy goals?

To handle this challenge, we consider two separate tasks:

- Classifying the policy theme of each campaign text: A primary dimension of analysis is policy issues, towards which the political parties express their positions and plans, and voters assess their alignment to these and express their views via votes. We develop a policy analysis model based on the popular label schema from the Comparative Manifesto Project (CMP) [Volkens et al. (2011)].

- Uncovering the rhetorical functions of campaign text: Knowing that a sentence discusses *increasing welfare funding* is not sufficient, because without its context it is hard to interpret how it adds to the policy proposal. For instance, a party may boast about its past achievements or talk about its proposed future action, which is important for characterising the party's position during a given election. Towards this end, we propose a speech act classification task, with the goal of decoding the context of discussion in a political campaign discourse. We develop the task definition by combining traditional speech act

classes [Searle (1976); Austin (1962)], which cover general speaker actions, with other domain-specific classes which capture more nuanced rhetorical functions commonly seen in electoral competitions.

In order to explain the details that can be extracted from campaign text, an example snippet taken from the election manifesto released by the Liberal Party of Australia, combining the output of both the analysis is given below:

> (1) Australia is now more vulnerable to economic shocks due to excessive deficits and debt spending. (2) We will get the budget back under control and within a decade, the budget surplus will be 1 per cent of GDP.

Though both sentences discuss the **Economy** policy issue, it can be seen that based on the speech act their rhetorical function varies — (1) *belief* and (2) *promise* respectively. First, the party acknowledges economic issues caused by excessive spending (1), and then promises a goal (2), built upon its beliefs and concerns, towards which the party is expected to function.

RQ2: The core of representative democracy is political parties making promises and delivering on them after they form government. But parties also use language to their advantage to not just make concrete promises. So an essential part of study is: how can we automatically detect verifiable promises, and ultimately build a system which can keep track of progress made towards election promises after government is formed?

We work towards this research goal in two parts.

- *Promise specificity prediction*: As part of our previous research question on rhetorical function classification, we categorize a sentence as a promise, or other speech act type. But we do not know the specificity of the promise, which can help to understand what information is present or not present. This is essential for studying party priorities, issue clarity, and political agenda, as well as a critical component of our goal of tracking progress made towards election promises [Pomper and Lederman (1980)]: it is obvious that a specific pledge is more verifiable than a vague pledge. For example, if a party says

  > *We will improve the livelihood of everyone*

  we can infer that this is a vague promise and cannot be easily tracked. At the same time a different promise such as

FIGURE 1.3: Guardian article on the budget decision related to ABC cuts, which indicates the 2013 election promise made by the Liberal Party of Australia has been broken.

No cuts to Medicare

is more specific and can be verified to improve party accountability.

- *Promise progress tracker*: After obtaining verifiable promises, the ruling party's progress towards these promises has to be tracked. This is non-trivial because there is no single source to check for progress updates, with possible sources including media-releases, news feeds, budget documents, Parliament or Congress bills, and other official documents (e.g., President's executive orders or interviews). This makes the task of identifying relevant evidence documents, and the subsequent task of progress status classification, a challenging one. An example evidence document expressing progress (which is **broken**) towards the promise *no cuts to Australian Broadcasting Corporation (ABC)* made by the Liberal Party of Australia during 2013 elections is shown in Figure 1.3.

RQ3: How can we automatically predict the popularity of petitions? Popular petitions can potentially influence policy changes. For example, a petition from *We The People* on

Making unlocking cell phones legal

got more than 110k signatures, and the petition led to the passing of legislation to restore the Digital Millennium Copyright Act in the US Congress, and was signed into law by the US President.[4] But, most petitions receive very few signatures. Automatically forecasting petition popularity can provide an opportunity at the time of submission for its authors to rewrite the text in a more persuasive fashion, as well for the petition team to adopt suitable campaign mechanisms to increase the reach of petitions, for example the decision to promote on the homepage of a petition website. Secondly, what are the reasons for a petition's popularity? This can help

[4] https://obamawhitehouse.archives.gov/blog/2015/07/28/how-we-are-changing-way-we-respond-petitions

the petition committee (or moderators) to use their resources optimally in order to engage with petition authors, with the goal of increasing the overall engagement of voters.

For this challenge, our primary focus is to predict the popularity of petitions using their textual content, which has been less explored in general, and not been the target of research in the political petitions context. For example, Elnoshokaty et al. (2016) investigated the use of hand-crafted features extracted from text in predicting popularity of Change.org petitions. Motivated by this work, since deep learning models generally perform better than simpler models using hand-crafted features only, we analyze the dependency between hidden representations learned by deep learning models and hand-crafted features. This can help to understand what is captured by deep learning models, which can in turn help the petition authors to improve its reach.

Though the set of research questions is addressed independently in this thesis, they can naturally combine together to produce end-to-end insights, as discussed in Chapter 4. Our contributions towards these research questions are detailed in Section 1.3.

1.3 Contributions

The contributions of this thesis are summarised as follows, according to the three research questions discussed in Section 1.2.

For the first research question, which is to analyze election campaign text based on policy position (Section 3.1) and speech acts (Section 3.2), we propose deep learning approaches which we train in supervised and semi-supervised scenarios. For sentence-level policy position analysis we use the label schema and annotations from Comparative Manifesto Project [Volkens et al. (2011)], which contains election manifestos covering many countries, written in different languages. It is a compelling source of data to demonstrate the utility of multi-lingual deep learning approaches. Common downstream use for these sentence-level annotations is in quantifying party ideological positions on a continuous left-right scale. For example, if a party discusses more about *economic* policies, then it tends to be right-leaning, and conversely if it discusses more *welfare* policies, then it tends to be left-leaning. Almost all prior work has focused on the sentence- and party-level tasks separately, though there is a natural dependency between them. Hence, our first work on policy theme classification, which was published at ALTA 2017 (Section 3.1), targeted the multi-lingual setting, and modeled both the sentence-level classification and document-level ideology regression tasks together using average word embeddings.

After that, we evaluated the use of sequential representations for both sentences and documents, again in a multi-lingual setting. We extended the model to a semi-supervised learning setting using a multi-task objective, where supervision of one task can compensate for another. This work was published at NAACL-HLT 2018 (Section 3.1). Following the policy position analysis work, we targeted the task of rhetorical function classification, for which we developed a label schema combining traditional classes with custom ones for the political domain. We also curated and annotated a dataset using media-releases and speech transcripts released by Australian political parties. Based on our observations in previous work, we used sequential modeling of sentences for this task with contextual word embeddings [Peters et al. (2018)], and proposed a semi-supervised extension of the approach. This work was published at *SEM 2019 (Section 3.2), co-located with NAACL-HLT 2019.

For the second research question towards improving political accountability, we analyzed election manifestos, to first filter out promises and classify their specificity (Section 3.3). Similar to our previous work, we modeled sentences using contextual word embeddings and sequential modeling of tokens, and posed it as an ordinal regression task, based on the class schema proposed by political scientists [Pomper and Lederman (1980)]. We annotated a dataset based on Australian election manifestos. We once again proposed a semi-supervised extension, and published the work at EMNLP-IJCNLP 2019. After promises are filtered out, a natural and challenging extension is to track progress after government is formed. For this we curated a dataset published by news organizations which manually track promises. We combined datasets from Australia, Canada, UK, and US political settings, and evaluated various approaches to serve as baselines for future work. This work is under submission to Computational Linguistics (Section 3.4).

Lastly, our third research question deals with analyzing text from voters, as distinct from the other work which deals with text from political parties. We targeted the petition popularity prediction task, using a dataset of UK and US government petitions. We use sequential modeling of petition text for popularity prediction, and show that it performs better than using hand-crafted features. This work was published at ACL 2018 (Section 3.5).

1.4 Structure

The thesis is structured as follows. Chapter 2 provides the background needed to understand the thesis. Chapter 3 enumerates the research outcomes in the form of several publications covering all five tasks: (a) manifesto policy theme classification and ideology prediction, (b) speech act classification to uncover the rhetorical roles in campaign text, (c) election campaign promise specificity prediction, (d) election promise progress tracker, and (e) government petition popularity prediction. Aside from the task differences, we group the work in Chapter 3 into three categories: (a) and (b) are both campaign text analysis; (c) and (d) target election promise analysis and progress tracking; and (e) focuses on voters advocacy analysis. The three categories of work are related to analyzing text from political parties and voters, with the common goal of improving transparency, awareness, and thereby trust in democracy. These similarities are discussed further in Chapter 4. This chapter also concludes the thesis with a recapitulation of research questions, analysis of research work, commonalities and gaps to deliver on end-to-end analysis, and future research directions.

Chapter 2

Background

'Text-as-data', as referred to by political scientists, deals with a broad set of techniques and approaches for analyzing the text generated by various political actors, to answer various questions in social science. It is essential because language is the primary medium of politics [Grimmer and Stewart (2013)], used by practically all stakeholders in representative democracies. During election campaigns, individual candidates and political parties express their policy positions through speeches, debates and manifestos. Speeches of other constitutional representatives such as the Governor-General or Queen provide the government with an opportunity to highlight its priorities for the future period [John and Jennings (2010); Dowding et al. (2010)]. Elected representatives debate legislation and vote for bills. After laws are passed, bureaucrats ask for comments before they issue regulations. The public discuss and express their views through social media, and they also launch campaigns for the change they wish to see, say through online petitions. Countries debate their stand on key issues in the United Nations. They regularly express statements that signal their priorities and also describe their motivations for cross-country ties. News reports publish the day-to-day affairs of regional and international issues, which also cover information from many other sources. Lastly, the adoption of social media and its use for widespread dissemination of political facts and opinion has changed the dynamics of political coordination and discourse. In all these cases, language, through spoken or written form, is central, and forms the primary documentary record.

Text has been a primary data source for political scientists [Gilardi and Wüest (2018)], e.g., Monroe and Schrodt (2008) posit that "text is arguably the most pervasive—and certainly the most persistent—artifact of political behavior". In addition to the traditional text sources, with

the rise of internet, there has been an exponential growth in web content — both participatory and collaborative. Due to this, political scientists have long recognized the need for automated text analysis. Given the cost and effort of analyzing even moderately sized collections of text, text-as-data approaches are becoming prevalent in political science [Gilardi and Wüest (2018); Grimmer and Stewart (2013)]. At the same time, it is also a keen area of focus by NLP researchers, studied under the banner of computational social science [Glavaš et al. (2019)].

In this thesis, we focus on improving transparency and voters' trust in representative democracy using NLP techniques. Specifically, we focus on *political parties* who reach out to voters during elections with their achievements and agenda, and *voters* who appraise based on those claims and promises. Transparency can influence the voters' trust in government, and in-turn the effective implementation of policies [OECD (2013)]. The goal of this chapter is to offer an overview of existing political text analyses, addressed by both the political science and NLP communities, covering the various text sources generated by *political parties* and *voters*, which provide the basic motivation for the research questions in this thesis. First we provide an overview of policy process in representative democracy, which details the process that leads to the creation of the different text sources relevant to the thesis (Section 2.1). In Section 2.2 we describe several sources of text available for analysis, and review applications built using these datasets. In Section 2.3 we discuss the major NLP challenges in relation to the datasets and target applications. Finally, in Section 2.4 we provide a summary of datasets, applications, and challenges which provide the technical motivation for this work.

2.1 Policy process

The ways to elect policy-makers and the decision-making processes can vary across countries. For instance, in the US, people select members of both the houses, while in the UK people select the lower house members, and the upper house members are appointed. The decision-making process also varies across democracies, in terms of the subtleties involved in how the bills are proposed and passed in the respective legislation processes. For example, the US President can issue directives via executive orders, which are seen as a way to achieve results that fail to get through Congress [Deering and Maltzman (1999)].

Though the policy-making process can be fluid in practise, a useful and commonly adapted sequence involves two major phases — policy proposal and policy-making. In turn these phases

can be sub-divided into more detailed stages based on the actions involved. This process whereby political parties formulate their proposals and request for a mandate to implement those actions is core to the idea of a representative democracy [Ashiagbor (2013)]. Among various political science tasks, policy analysis is one topic where the use of NLP is relatively rare [Gilardi and Wüest (2018)].

We provide a general overview of the two major phases and the stages involved in each, which applies to all the democracies and we point out specific cases wherever applicable. After that we discuss voter advocacy, which allows voters to participate more actively in the process.

Policy proposal

Citizens have requirements that they want the governments to address. Political parties aggregate these demands from voters and other groups, and articulate public policy options as a response. Policy drafting involves different tasks: identifying and prioritizing issues for policy attention; analyzing policy options; developing policy proposal documents; reviewing the policy proposal with the help of experts; and finalizing the policy proposal. The proposed policy options are usually based on the national interest, party ideology, and those that can satisfy the party goals (e.g., winnability). It is also the responsibility of the political parties to raise awareness about policy issues that may otherwise be neglected. Elections, held at periodic intervals, provide voters with the opportunity to rate and choose among political parties offering different policy proposals [Ashiagbor (2013)]. The cycle of elections vary across countries, e.g., 4 years for UK government and US Presidential elections; 5 years for Canadian federal and France Presidential elections; and 6 years for Finland Presidential election. Campaign communications can influence a party's reputation, credibility, and competence, which are primary factors in voter decision making [Fernandez-Vazquez (2014b)]. Hence the political parties need to ensure that the proposed public policy sets the priorities for government action.

Policy making

Once government is formed, parties require strategies to convert their policy proposals into functioning processes [Ashiagbor (2013)]. Depending on the system, policies are implemented either through the executive and legislative institutions, or the administrative groups they influence. Opposition parties evaluate and present alternatives to ruling party proposals. In subsequent elections, citizens can use their votes to appraise parties based on their actions. In addition to this power, citizens expect transparency related to the government's actions. Hence, political parties, ministers, and officials inform the public of their actions in the Parliament through systems such as Hansard and other channels.

At its simplest level, policy-making can be understood as a sequence of four stages: agenda setting, formulation, implementation, and evaluation [Shiffman (2016)].

Agenda setting refers to the stage of a problem being initially sensed by policy actors, and possible solutions being put forward. This can be expressed using the Governor-General or Queen's speech, media-releases, Parliament motions, parliament members' questions, etc. In many political systems, the chief of state delivers a formal statement of the proposed legislative programme, which sets the priorities of the government. In Australia, the Governor-General, who is the Queen's representative, gives the speech at several occasions including the opening of a new parliament and the appointment of a new Senate [Dowding et al. (2010)]. In Britain, the Queen's Speech details the legislative measures intended to be enacted in the next Parliament session [John and Jennings (2010)]. The government and its corresponding administrators for policy issues provide updates through media-releases and speeches. Parliament activities can be tracked using motions, question time, etc.

Formulation stage deals with the development of specific policy options by excluding infeasible options and favoring some solutions over others. It can be done by individual ministers or a cabinet team and also supported by relevant administration teams. Among the shortlisted plans, one policy is accepted by the decision-makers. In many cases, a policy is adopted when Parliament (or Congress) passes a law, or through a President's executive order or Supreme Court ruling. This can result in federal laws, federal decrees, ordinances or directives.

Implementation: The government implements their decisions using public administration tools. Monitoring the implementation of measures is a key task for the relevant ministry.

Evaluation: The results of policies are evaluated by both state and societal actors, often leading to the re-conceptualization of policy issues and solutions in the light of feedback obtained in relation to the policy in question, possibly leading to the start of a new iteration of the cycle [Howlett (2009)].

These phases are critical in any representative democracy, and the core idea is to link citizens to the functioning of representatives, as also summarized by the responsible party doctrine [Wilson (2002)],

> *Parties should act in distinct organizations, in accordance with avowed principles, under easily recognized leaders, in order that the voters might declare by their ballots, not only their condemnation of any past policy, by withdrawing all support from the party responsible for it; but also and particularly their will as to future administration of the government, by bringing into power a party pledged to the adoption of an acceptable policy.*

Other than the involvement of the public via voting, it is necessary to maintain transparency of process, and thereby earn the trust of voters. This can be achieved by seeking feedback on policies from the public [Gilardi and Wüest (2018)], both at the policy proposal and policy-making stages. For example, in the lead up to the 2012 US Presidential elections, the Democratic Party conducted a survey, by listing policy priorities of the Republican Party, and asking their supporters to identify the policies which concerned them the most. They were also asked to rank policy goals such as unemployment and immigration reforms. Similarly, the Conservative Party of UK launched a website called "stand Up, speak Up", to ask for feedback from voters, on its policy proposals for the 2010 elections [Ashiagbor (2013)]. After government is formed, social media enables the voters to provide feedback on policy agendas and implementation. Among social media channels, e-petitions provide a unique opportunity to raise issues bottom-up from the voters. Through voter support for petitions, those issues can gain the attention of parliamentarians, which can influence the agenda setting of policy-makers [Böhle and Riehm (2013)].

2.2 Artefacts

The policy process discussed above (Section 2.1), generates several sources of information, pertaining to the analysis of the democratic roles of political parties and voters. In this section we enumerate those text sources, and mention any existing popular projects and annotations built on those sources including the label schema with example annotated text. Data behind the policy proposal and policy-making processes (see Section 2.2.1 and 2.2.2) is generated by politicians and political parties, while petitions are written by citizens and responded to by governments (as we will discuss in Section 2.2.3). News can be seen as a way to connect different stakeholders in a democracy — through disseminating politically relevant information to citizens, thereby acting as a public watchdog; and providing a representative platform to help the voices of the public be heard [Müller (2014)] (discussed further in Section 2.2.4). These form the core artefacts for our work, and we also discuss other sources which are relevant and can be explored for future research, such as social media, which is relevant to all the stakeholders in a democracy, and provides a platform to engage in participatory and collaborative roles, which consequently enable the tracking of day-to-day political events in a faster way than traditional media. We provide an overview of social media data sources in Section 2.2.5, as they have also been a popular source of text for much NLP work.

2.2.1 Election campaign text

Among a myriad of data sources, election campaign text is a core artefact in political analysis. Using campaign communication, the political parties express their opinion, stance and performative actions on different policy issues. Based on this, they appeal to the general public to garner their support in elections. They use campaign communication especially to place themselves ahead of other parties in the electoral race, by detailing their past achievements, criticizing opposite party policies and actions, and by emphasizing certain issues. Hence the campaign communication can influence a party's reputation, credibility, and competence [Fernandez-Vazquez (2014a)]. There are different media for conducting election campaign — press releases, speeches, manifestos, social media, debates, interviews, etc., which each have different styles of expression. Press releases are typically a set of documents released over an election cycle, where each document focuses on a single policy issue and the proposed action (or its updates) from a party. Speeches are made by leaders of a party over time, across different locations, for

example, different states in the US, and usually tailored to its audience group, the maritime industry for instance.[1] A single manifesto, which is a large compilation of party views and goals, is released by each party, for each election cycle. Debates and interviews involve the mass media — to question, discuss and differentiate between opponents' positions; and to question and understand contestants' (or parties') positions, respectively.

First we consider digitized collection of election manifestos, which are a traditional data source used to analyze parties' policy goals (Section 2.2.1.1). We use this dataset in our work not only because it is commonly used by political scientists, but also due to its longitudinal availability across countries. Then we discuss other data sources such as media releases, presidential debates, advertisements, and speeches (in Section 2.2.1.2). Compared to manifestos, these sources do not have large scale data over time covering many geographic regions, for many reasons — debates are a relatively recent mechanism (for example, Australia has had debates since 1984) and are held only in a handful of countries;[2] for most countries, there is no central digitized collection of speeches, media-releases and interviews for political parties; and typically when there is a change in leadership, the documents prior to that are archived, for example, the Liberal Party of Australia's website[3] only includes documents created in 2019 or later. Despite the patchy nature of the data collections, they are still a valuable source of information for many fine-grained analyses to capture the campaign dynamics in a particular election cycle, which maybe influenced by various context factors [Laver et al. (2003)]. We use media-releases and speech transcripts in the Australian political setting for campaign rhetorical functions analysis.

2.2.1.1 Manifestos

Manifestos provided by political parties help to streamline the election campaign — they make the job of candidates easier in prioritizing policy goals in their campaign. Manifestos are published dutifully by the political parties in order to document a summary of their party positions, which act as their blueprint for future government. They are disseminated to candidates and supporters in a top-down fashion, and also attract media attention to party positions. This is distinct from other data sources such as debates or media-releases which can contain more than one document released during an election cycle, and can be related to a particular individual's portfolio. The importance of election manifestos is evident from the attention paid by political

[1] https://anthonyalbanese.com.au/address-to-maritime-industry-australia-ltd-forum-promot

[2] https://www.theguardian.com/commentisfree/2010/apr/07/debates-leaders-media

[3] https://www.liberal.org.au/articles

CMP Coding Scheme
1: External Relations
2: Freedom and Democracy
3: Political System
4: Economy
5: Welfare and Quality of Life
6: Fabric of Society
7: Social Groups

TABLE 2.1: Comparative Manifesto Project — major policy themes

scientists. One major reason is their primary function — providing a concise collection of valid party positions — thereby establishing superiority over all other campaign text sources [Eder et al. (2017)]. Another larger reason for manifestos' importance, is due to the regularity with which parties produce these documents, and the effort of political scientists through projects such as the Comparative Manifesto Project (CMP) to digitize and provide usable data [Volkens et al. (2011)].

CMP is one of the most widely used datasets by political scientists, and contains 4492 manifestos in various languages, covering over 1000 parties across 56 countries, from elections dating back to 1920. The amount of text is not uniform across all the countries, and the details are summarized in Table 2.2. The goal of CMP is to provide a large collection of data to support studies on electoral processes [Zirn et al. (2016)]. Many of the manifestos have been manually annotated at the sub-sentence level with one of over fifty fine-grained policy themes,[4] divided into 7 coarse-grained topics (see Table 2.1). For example, *Foreign Special Relationships: Positive* and *Foreign Special Relationships: Negative* are fine-grained policy themes under the major category of *External Relations*. Political scientists who are native language speakers, are usually involved in annotating manifestos in the respective countries. They go through a training process to ensure a consistent understanding of the label schema across countries [Lacewell and Werner (2013)]. Annotators split the manifesto text into sub-sentences (or so-called 'quasi-sentences'), each of which contains exactly one message or policy theme [Werner et al. (2011); Merz et al. (2016)].

The first version of the label schema was designed by Robertson (1976) to analyse the dynamics of party competition in Britain. It was extended by CMP, with extensive and well-designed policy themes, and is used to study policies across many countries [Klingemann et al. (1994);

[4] https://manifesto-project.wzb.eu/coding_schemes/mp_v5

Country	Start year	#Manifestos	#Sentences	#Parties
Albania	1991	38	8675	11
Armenia	1995	28	5327	16
Australia	1946	111	39731	11
Austria	1949	83	49530	10
Azerbaijan	1995	9	955	6
Belarus	1995	8	979	8
Belgium	1946	177	187035	26
Bosnia-Herzegovina	1990	57	21358	20
Bulgaria	1990	59	25252	27
Canada	1945	94	41518	9
Croatia	1990	73	30589	38
Cyprus	1996	30	24720	11
Czech Republic	1990	63	34115	25
Denmark	1945	235	29893	19
Estonia	1992	47	17291	22
Finland	1945	153	28079	13
France	1946	116	26556	23
Georgia	1990	54	12063	40
German Democratic Republic	1990	15	2757	15
Germany	1949	89	74606	19
Greece	1974	84	76756	17
Hungary	1990	43	45829	13
Iceland	1946	117	20742	19
Ireland	1948	103	55505	17
Israel	1949	198	25611	55
Italy	1946	178	100853	52
Japan	1960	123	23550	25
Latvia	1993	61	6967	36
Lithuania	1992	52	40485	26
Luxembourg	1945	75	62346	9
Malta	1996	4	4018	2
Mexico	1946	94	69212	23
Moldova	1994	28	6187	15
Montenegro	1990	59	18852	29
Netherlands	1946	178	183720	28
New Zealand	1946	101	72044	11
North Macedonia	1990	74	73767	28
Northern Ireland	1921	33	2589	3
Norway	1945	130	167233	11
Poland	1991	60	22182	33
Portugal	1975	101	103718	17
Romania	1990	52	15585	30
Russia	1993	45	12545	24
Serbia	1990	80	31459	39
Slovakia	1990	69	37501	29
Slovenia	1990	63	37973	22
South Africa	1994	17	6423	6
South Korea	1992	28	24949	16
Spain	1977	144	217662	33
Sri Lanka	1947	13	2324	2
Sweden	1944	137	36118	9
Switzerland	1947	146	42242	17
Turkey	1950	77	87633	23
Ukraine	1994	45	5702	30
United Kingdom	1945	88	57993	13
United States	1920	53	38111	5

TABLE 2.2: Statistics of CMP dataset, 2019 version, showing the *start year* (or election) from which manifestos are available for each country (all continuing to the present day), number of manifestos, coded sentences and parties.

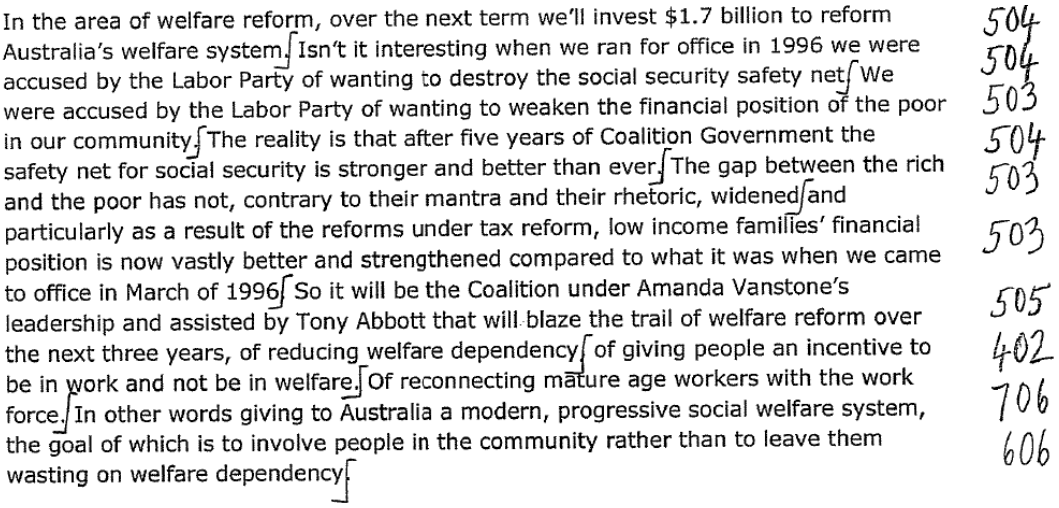


FIGURE 2.1: Manifesto snippet taken from the Comparative Manifesto Project for the Liberal Party of Australia, 2001. Manifestos from older elections are made available by scanning the corresponding printed copies, as shown here. Digitized text and its annotations are provided by the project — ∫ denotes a coherent segment. The starting digit of the fine-grained policy theme denotes its major policy category's code (Table 2.1). See text for code description.

Volkens et al. (2011)]. An example snippet from the Liberal Party of Australia 2001 election manifesto is given in Figure 2.1. A description of fine-grained themes from the snippet is as follows:

(a) 402 Incentives: Positive (under the theme of *Economy*)

(b) 503 Equality: Positive (under the theme of *Welfare and Quality of Life*)

(c) 504 Welfare State Expansion (under the theme of *Welfare and Quality of Life*)

(d) 505 Welfare State Limitation (under the theme of *Welfare and Quality of Life*)

(e) 606 Civic Mindedness: Positive (under the theme of *Fabric of Society*), and

(f) 706 Non-economic Demographic Groups (under the theme of *Social Groups*).

This can be seen as fine-grained policy topical analysis. Other than the facility to download the required manifestos from its online repository,[5] the corpus is stored in a standardized format in a database, and can be accessed easier via the R package *manifestoR*[6] [Merz et al. (2016)]. More details about the dataset and the complete label schema are provided in Section 3.1.

[5] https://manifesto-project.wzb.eu/
[6] https://github.com/ManifestoProject/manifestoR

Left	Right
103 Anti-Imperialism	104 Military: Positive
105 Military: Negative	201 Freedom and Human Rights
106 Peace	203 Constitutionalism: Positive
107 Internationalism: Positive	305 Political Authority
202 Democracy	401 Free Market Economy
403 Market Regulation	402 Incentives: Positive
404 Economic Planning	407 Protectionism: Negative
406 Protectionism: Positive	414 Economic Orthodoxy
412 Controlled Economy	505 Welfare State Limitation
413 Nationalisation	601 National Way of Life: Positive
504 Welfare State Expansion	603 Traditional Morality: Positive
506 Education Expansion	605 Law and Order: Positive
701 Labour Groups: Positive	606 Civic Mindedness: Positive

TABLE 2.3: 26 out of 57 policy themes coded as *left* and *right* by political scientists. The starting digit of the fine-grained policy theme denotes its major policy's category code (see Table 2.1).

Party position indicators

In addition to the policy theme, party position analysis is an important task for political scientists. One important theme deals with studying party positions on the left–right ideological scale, for which the fine-grained policy themes are placed on left-right dimension by political scientists [Laver and Budge (1993)], as shown in Table 2.3. For instance, a *negative* mention of the *military* topic is categorized as *left*. The remaining 31 out 57 policy themes are considered *neutral*. Other than the sentence-level labels, the manifesto text also has a document-level score that quantifies the party's position on the left–right spectrum in that particular election. Different approaches have been proposed to derive this score, based on alternate definitions of "Right–Left" [Slapin and Proksch (2008); Benoit and Laver (2007); Lo et al. (2013)]. Among these, the Left-Right (RILE) scale is the most widely adopted [Jou and Dalton (2017); Merz et al. (2016)], and Lowe et al. (2011) showed that the index correlates highly with popular expert survey scores for importance estimates from Benoit and Laver (2006), across several policy dimensions. RILE is defined as the difference between *right* and *left* positions on pre-determined policy themes (shown in Table 2.3) across sub-sentences in a manifesto. Apart from RILE, expert survey scores are gaining popularity as a means of capturing party's ideology score. In Benoit and Laver (2007), the authors show that the expert surveys are less noisy than the RILE

estimation from CMP, and they also allow placing different weights on the specific policy dimensions which constitute left and right from one country to the next. Although expert surveys are more subjective than RILE, they are considered to be the gold-standard in political science. Along the same lines, the Chapel Hill Expert Survey (CHES) [Bakker et al. (2015)] is one such popular survey, which comprises aggregated expert surveys on the ideological position of various European political parties, available longitudinally since 1999. During each iteration of the survey, starting from 1999, the number of parties, questions and countries covered has increased — the most recent survey includes 31 countries and 268 parties.[7] For Australia, an analogous resource is the Australian Election Study [Cameron and McAllister (2019)].

Other than general left-right analysis, the surveys contain party position indicators on finer dimensions — namely social left-right and economic left-right. Apart from ideological position, the surveys also include issue salience of parties: how much importance the party gives to different policy issues [Benoit and Laver (2006); Bakker et al. (2015)]. Similarly, the salience score can also be computed based on public surveys [Cameron and McAllister (2019)].

Political manifesto text analysis is a relatively novel application, at the intersection of Political Science and NLP. One line of work in this space has been on sentence-level (or sub-sentence level) classification, including classifying each sentence according to its major political theme (using the seven categories in Table 2.1) [Zirn et al. (2016); Glavaš et al. (2017a)] and fine-grained policy themes [Verberne et al. (2014); Biessmann (2016)]. Zirn et al. (2016) developed a global model to capture the topic shift between sentences, in addition to the sentence-level topic classifier based on a Markov Logic Network. They show that a global model which also considers shifts between sentences provides the best performance. Glavaš et al. (2017a) handled manifestos spanning multiple countries and languages, using multilingual word embeddings obtained by mapping monolingual embeddings to a common space using the linear translation approach [Mikolov et al. (2013)]. Biessmann (2016) developed a fine-grained policy theme classification model for German manifestos using a Logistic Regression classifier and bag-of-words representation. Verberne et al. (2014) targeted fine-grained classification for Dutch manifestos with a custom label schema (more than 200 categories) proposed by Lipschits (1977), using Support Vector Machines and Winnow classifiers. At the document level, there has been work on using label count aggregation of (manually annotated) fine-grained policy positions, as features for inductive analysis [Lowe et al. (2011); Däubler and Benoit (2017)]. Lowe et al. (2011)

[7] Sample questions can be found here: https://www.chesdata.eu/s/Belgium_Questionnaire_2014.pdf

examined the popular RILE scaling, proposed an alternative based on the logarithmic odds-ratio, and evaluated using annotated manifestos. Däubler and Benoit (2017) applied a Bayesian item-response theory model to policy-theme counts from annotated manifestos, where they treat the policy-themes as "items" and ideological positions as latent variable. Text-based approaches have used dictionary-based supervised methods and unsupervised factor analysis [Bruinsma and Gemenis (2019); Hjorth et al. (2015)]. Bruinsma and Gemenis (2019) use the Wordscores approach proposed by Laver et al. (2003), which uses a set of reference documents to scale a new set of documents based on word-occurrences in text. Scores for unseen documents are estimated as the average of the scores of the words contained in them. In turn the words' scores are computed using the reference texts in which the word type is present, and by averaging these documents' scores weighed by the relative frequency of the tokens. Hjorth et al. (2015) compared Wordscores and Wordfish [Slapin and Proksch (2008)] using German and Danish manifestos. Unlike Wordscores, Wordfish does not use any reference text, but fits the term-frequency matrix of a corpus to a standard Poisson count model to infer the (latent) ideology position. Hjorth et al. (2015) showed that Wordscores can produce outputs that are closer to the ground-truth, and Wordfish requires a larger corpus to model the word usage differences.

Apart from policy-theme-based sentence and ideology-based document-level tasks, there are other semantic tasks such as finding agreement and disagreement between parties on coarse-level policy issues [Menini et al. (2017)]. They employed keyphrase extraction and clustering to align sentences discussing the same topic. For keyphrase extraction, they used the rule-based Keyphrase Digger approach [Moretti et al. (2015)]. To classify the agreement between pairs of statements, they use supervised approaches with lexical, surface and semantic features [Menini and Tonelli (2016)]. Election manifestos have been analyzed for prospective and retrospective messages [Müller (2018)]. Retrospective statements are used to evaluate parties' position and actions of the present and past, and prospective statements refer to actions for the future. The authors studied 576 manifestos across nine democracies, and found that the ideologically extreme parties give lesser emphasis towards describing future actions than mainstream parties. Opposition parties, irrespective of ideology, convey the past and present (retrospective) statements more negatively.

Labor has listed columns with zero, zero, zero and zero allocated, for ending the Medicare freeze in their costings. The Morrison Government has ended the freeze which Labor started. We have invested $1.6 billion in primary care, working with GPs to deliver more flexible options for care and to keep patients healthy. People should remember that Bill Shorten, as Assistant Treasurer stopped listing medicines because he couldn't manage the budget. And Bill Shorten has already cut $115 million from the private health rebate which will hit pensioners and low income earners from 1 July 2019 and require them to hand money back. By managing the economy and the budget we can deliver the important health services that Australians need.

TABLE 2.4: Snippet from 2019 election campaign — The Liberal Party of Australia's media-release

2.2.1.2 Other campaign sources

Apart from manifestos, political parties also use media-releases, and speeches to convey their views and opinions (e.g., the Liberal Party of Australia's articles[8]). Media-releases are typically a summary of party position on a particular issue, and released either by the party or members with portfolios relevant to that policy issue. During the course of the election campaign, leaders make speeches at different locations, which often contain party goals delivered in a style to persuade the targeted audience. Though they are monologue in style, from the example media-release shown in Table 2.4, we can see that they express various rhetorical functions — criticizing the opposite party's policy position and contrasting this with their past achievements and future goals — and are also dynamic over time within an election cycle, unlike manifestos. In addition, televised public debates have become a focal point of election campaigns, for example, American election debates. They are also the most followed, viewed, and the most researched political television programs. The American Presidency project at the University of California, Santa Barbara [Woolley and Peters (2011)], includes all the presidential documents — election debate transcripts, White House press briefings, proclamations, and executive actions. Presidential debates have been intensively studied from the Nixon–Kennedy contest of 1960 election to the present [Isotalus (2011)]. A snippet from 2016 US presidential debate between Donald

[8] https://www.liberal.org.au/articles

WALLACE: Mr. Trump, you're pro-life. But I want to ask you specifically: Do you want the court, including the justices that you will name, to overturn Roe v. Wade, which includes—in fact, states—a woman's right to abortion?

TRUMP: Well, if that would happen, because I am pro-life, and I will be appointing pro-life judges, I would think that that will go back to the individual states.

WALLACE: But I'm asking you specifically. Would you like to...

TRUMP: If they overturned it, it will go back to the states.

WALLACE: Secretary Clinton?

CLINTON: Well, I strongly support Roe v. Wade, which guarantees a constitutional right to a woman to make the most intimate, most difficult, in many cases, decisions about her health care that one can imagine. And in this case, it's not only about Roe v. Wade. It is about what's happening right now in America.

TABLE 2.5: Snippet from 2016 US President debate at the University of Nevada, Las Vegas

Trump and Hillary Clinton is given in Table 2.5. Debate text provides a topically-aligned dialog between candidates, which can be used to compare their positions on election issues.

Benoit (2001) studied the functional theory of campaign discourse, applied to presidential television spots from 1952–2000. Because elections can be seen as a competition, candidates tend to praise their own strengths and to attack their opponents' weaknesses. When subjected to such attacks, candidates may also choose to refute those accusations. These three functions can occur on the grounds of both policy and character. This work compares television spots of parties through the functions of utterances. Presidential campaign text have also been studied for the relationship of the topic of discourse—the proportion of utterances focusing on policy and character—and election outcome. Benoit (2003) studied a large sample of presidential media, spanning 1948–2000, and show that presidential candidates who discuss policy more frequently (and character less) than their opponents are more likely to win elections. Kleinnijenhuis and De Nooy (2013) used autoregressive models to analyze adjustment of issue positions based on news items, during the 2016 Dutch election campaign. Party positions on different policy issues are obtained from daily content analysis of newspaper and television news. They show that parties adjust their issue positions and explain it based on inter-party relations, ideological reciprocity, and realignment. Campaign advertisements produced by candidates in the 1976-1996 US presidential elections were modeled using Logistic Regression in order to understand

the dynamics of issue ownership [Damore (2004)], where the target variable indicates whether an advertisement discusses a *trespassed issue* or *party-owned issue*. Their results indicate that candidates' decisions to trespass into issues owned by their opposite party are a function of their competitive standing, partisanship, issue salience, and the tone of campaign messages. All these approaches were based on manual annotations of campaign data, whereas the goal of this thesis is to automate them using natural language processing techniques.

Gautrais et al. (2017) proposed an approach combining standard topic modeling with *signature* mining for analyzing recurrent topics in the campaign speeches of Clinton and Trump during the 2016 US Presidential elections. They showed that Trump was more diverse in the choice of his recurrent topics. Liu and Lei (2018) studied sentiment and topics using 2016 US Presidential election campaign text. They found that Hillary Clinton's discourse focused more on socio-economic topics, while Trump's discourse focused on trade- or business-related topics. They also found that Trump's speeches were more negative than Clinton's. Another important aspect of political campaign are the rhetorical functions, because the speakers at these events prioritize maintaining the crowd's attention [Strangert (2005)]. Heritage and Greatbatch (1986) manually analysed speeches from British political party conferences, associating each instance of applause with message types (e.g., external attacks, statements of approval), rhetorical devices (e.g. contrast, headline-punchline), and performance factors (e.g. speech stress or body language). They find most of these factors to be positively correlated with applause, while rhetorical devices explain the majority of applause instances. Gillick and Bamman (2018) proposed automatic modeling of applause in campaign speeches. They analyzed presidential campaign speeches, and used audio data to automatically extract and align labels for instances of audience applause. They show that lexical features carry the most information for applause prediction, in addition to audio signals.

For the analysis of policy proposal, also referred to as policy agendas during elections, the content of election manifestos, media-releases, speeches, political ads and other political communications are paramount. News and social media content are also valuable resources, which we discuss separately in Section 2.2.4 and Section 2.2.5. Using the election campaign text — manifestos, media releases and speech transcripts — we survey policy themes and discourse, and also evaluate the effectiveness of the approach based on its ability to model party ideology and issue salience indicators.

The "Concealed Carry Reciprocity Act," a top priority for the NRA and other gun-rights groups, passed 231-198. | AP Photo

House passes concealed carry gun bill in win for GOP and NRA

By **JOHN BRESNAHAN** | 12/06/2017 04:45 PM EST | Updated 12/06/2017 06:39 PM EST

Share on Facebook Share on Twitter

The House passed legislation to permit concealed carry license holders to conceal a handgun in other states, the first time Congress has taken action on a gun bill since President Donald Trump was sworn into office.

The "Concealed Carry Reciprocity Act," a top priority for the National Rifle Association and other gun-rights groups, passed 231-198. Six House Democrats crossed the aisle and voted for the measure, while 14 Republicans opposed it.

FIGURE 2.2: Snippet from Politico story on US Congress passing a bill to carry concealed gun.

2.2.2 Policy-making

After elections, laws result from a complex social process where again language is central to that process. We provide a short description of the artefacts related to the policy-making process here, which enables tracking the policies which successfully pass through all the stages.

An important source of information which can help understand the role of all the parties in governance is legislation bills (proposed laws) and voting records. Roll call votes in the US congress record how legislators vote on bills.[9] Other than voting, even more interesting are the subtleties involved in the legislative process, which includes debate among legislators [Laver et al. (2003); Quinn et al. (2006)]. A bill is formally proposed by a Congress member, aka sponsor. Once proposed, it is routed to a suitable Congressional committee. Members of that committee recommend bills for consideration and voting by the full chamber of members, both in the House of Representatives and the Senate. [Yano et al. (2012)].

Bills also contain rich metadata — e.g., sponsors of a bill, policy area and category (trivial, recurring, technical or important) — providing further insights into the functioning of parties

[9] https://www.govtrack.us/

on the spectrum of *constructive democracy*. Yano et al. (2012) studied bill survival in Congressional committees modeling its text and other meta features using a Logistic Regression model. They found that, in additional to bill content, its category, and sponsor's partisanship and committee member status influenced a bill's future. Similarly in the House of Commons of the UK Parliament, members of Parliament put forward their proposal (also known as debate's motion), and other members of the House debate and vote on the motion. The legislation process is also usually covered in press-releases, news (see Figure 2.2 for an example) and social media platforms [Kratzke (2017); Grimmer (2010)], which can enable public participation [Yates and Greenberg (2014)]. For the parliamentary debates, the Hansard is a valuable resource. It contains transcripts of parliamentary debates in Britain and many Commonwealth countries. For example, in Australia[10] it includes proceedings of the Senate, House of Representatives, the Federation Chamber[11] and all parliamentary committees.

Several political analyses have studied Congress voting data. One such effort is ConVote [Thomas et al. (2006)], a corpus of US congressional debate transcripts. It consists of 3,857 speeches organized into 53 debates. Each speech is tagged with the identity of the speaker and a "for" or "against" label derived from congressional voting records. The task is to predict the members' vote given their speech text. They used a SVM-based approach, and also leveraged the discourse structure in the debates by augmenting the model with same-speaker constraints and different speaker agreements. Burfoot et al. (2011) extended the content-only model with citation information based on speaker references in text, using a collective classification approach. They showed that collective approach performs better than content-only setting. Burford et al. (2015) evaluated an alternate setting, where the links between speeches are based on text similarity (or implicit links). They showed that using these implicit links boosts the performance of content-only classifier. A different formulation of the problem is to predict the legislators vote given only the bill text. This can be seen as modeling legislators directly. Kraft et al. (2016) used a simple average word-embedding approach to handle the task. Kornilova et al. (2018) extended it using a convolutional neural network model, and also incorporated additional metadata — bill sponsor details. They show that the proposed model which utilizes metadata performs best, not only for the same Congress session test-set (in-session), but also for the out-of-session setting. Iyyer et al. (2014) applied a recursive neural network to identify the political position evinced by a sentence. They extended the ConVote dataset with crowdsourced annotations at both phrase and sentence level in order to train the model. Bhatia and Deepak (2018) modeled

[10] https://www.aph.gov.au/parliamentary_business/hansard

[11] Parallel body to expedite Parliamentary business, and the decisions must be unanimous.

coarse-grained political ideology of the speech text as a function of sentiment polarities towards a set of topics. They evaluated their approach, to find topic-specific sentiments using an recurrent neural network-based approach and a Logistic Regression classifier to predict ideology, using the ConVote dataset.

In relation to other democracies, Abercrombie and Batista-Navarro (2018a) applied opinion mining methods to UK parliamentary debate transcripts to classify the sentiment of speakers as being either positive or negative towards the motions proposed in the debates. They automated the task using multi-layer perceptron, and also compared the classifier's performance using both manually annotated sentiment labels and the speakers' division votes ('aye' or 'no'). Abercrombie and Batista-Navarro (2018b) modeled opinion-topics using UK House of Common parliamentary debates. They evaluated the use of Public Whip's policies — policies represent both a policy topic and a polarised position towards it, e.g., Abortion, Embryology and Euthanasia *Against* — to train a classifier for inferring policy positions in UK parliamentary data. They showed near-perfect agreement with human annotation, but did not evaluate using machine learning approaches. Abercrombie et al. (2019) focus on the task of policy topic classification using UK parliamentary debates, where the label schema is adopted from the popular CMP (Section 2.2.1.1). They propose annotation guidelines for the two granularities in label schema, and evaluate various techniques to automate the task. Vilares and He (2017) proposed a Bayesian model to detect people's perspectives using debates from the House of Commons of the UK Parliament. The Bayesian approach considers topics (or propositions) and their associated perspectives (or viewpoints) as latent variables. Rheault (2016) developed an approach by adapting domain-specific corpora to automatically measure the levels of "anxiety" in Canadian House of Commons debates. In politics, "anxiety" is an important emotion because it is is tied to decision-making under uncertainty. They find that less familiar political issues, such as immigration, are more anxiogenic than average.

Kauffman et al. (2018) proposed approaches to predict legislator's alignment, in terms of whether they agree or not, with an organization that has taken a position on some legislation. They study the prediction of alignment scores between each member of the California state legislature and a select set of state-recognized organizations, using the assembly speeches. Duthie and Budzynska (2018) proposed to use a deep modular recurrent neural network approach for automatically extracting the ethos relation of the speaker towards a target (a single target entity, individual or group) from UK parliamentary debates. The ethos relations can be support or attack towards the target. They created a corpus of ethotic relations from the UK Hansard, by extracting speech

transcripts from 1979 to 1990. Other than using mentions to construct relational data from debates [Burfoot et al. (2011)], Salah et al. (2013) use agreement information to construct graphs from UK House of Commons debates. They represent each debate as a graph with speakers as nodes and information exchanges as links. Links are labeled based on the attitude of speakers, i.e., if both are negative or positive, then a link is labelled as *supporting*; otherwise it is labelled as *opposing*. Further detailed discussion of sentiment and position-taking analysis with parliamentary debates can be found in Abercrombie and Batista-Navarro (2020).

Oral and written questions play a prominent role in social interactions, and have specific rhetorical functions compared to other forms of informational exchange. For example, Question Time in the House of Representatives in the Australian Federal Parliament begins at 2pm on every sitting day, and is a period where members of the parliament can ask Ministers questions and receive a detailed response [Turpin (2012)]. Parliamentary party groups use parliamentary questions not only to obtain answers from the government, but also to further their own interests. Parties use parliamentary questions to retain their issue-ownership [Green-Pedersen (2010)], also to target members from other parties that they compete with for voters [Walter (2014)]. Zhang et al. (2017) propose an unsupervised technique to extract motifs that recur in questions, and group them based on their latent rhetorical role. They analyzed question session data in the UK parliament, and found that the resulting categorization captures salient aspects of political discourse, including the distinction in questioning behavior between government (uses more agreement type of questions) and opposition parties (uses more concede/accept and condemnatory type of questions). They also show that younger opposition members tend to ask more condemnatory questions compared to older members, who favor concede/accept questions. Overall the questions from representatives can be useful for analyzing policy progress, but are overall seen as symbolic due to their limited policy consequences [Van Aelst and Vliegenthart (2014)].

Other policy decisions are available through the respective minister's communication channels. For instance, Greg Hunt (Minister of Health in Australia since January 2017), publishes media-releases updating progress made in the *health* policy-area. One such example is a policy decision update dealing with diabetes patients getting free access to the FreeStyle Libre flash continuous glucose monitoring (CGM[12]) system through the Scott Morrison Government's *CGM initiative*, that was announced in the budget.

[12] http://shorturl.at/fmQ08

Budget text is a core artifact in itself, and it provides a summary or plan of the intended revenues and expenditures of that government. Budget items convey the expenditure-related priorities and decision choices of the legislators. Similar to parliamentary debates, budget sessions allow members to propose motions and amendments, and the House may vote on the amendment and/or the main motion. In addition to the budget numbers, expenses and revenues are also monitored through systems such as the Mid-Year Economic and Fiscal Outlook (MYEFO).[13] The MYEFO keeps tracks of all the expense- and revenue-related decisions made since the release of the budget, and revises the budget aggregates.

National-level policies are influenced by various local as well as international factors. One prominent resource to monitor international activities is the UN debates. Political scientists have observed that the UN influences state behavior, and relations can be observed in the national policies [Riggs (1967)]. A major UN debate collections is associated with the annual sessions of the United Nations General Assembly since 1947. It is a forum at which leaders present their country's perspective on major issues related to world politics. Thereby, the General Debate provides a unique source of information on the preferences of states on a wide set of policy issues which are of strategic importance for the world. The UN General Debate Corpus (UNGDC) [Jankin Mikhaylov et al. (2017)] introduces an English language corpus of 8,093 texts of debate statements from 1970 to 2018. As a second data source, the UN Security Council is one of the six principal organs of the United Nations. The Security Council meets regularly in a public format allowing for the global public to follow key debates and votes. Although public meetings do not provide the details in entirety, these meetings are among the best reported regular activities of the United Nations. The dataset from Schönfeld et al. (2019) is a collection of UN Security Council debates, covering 4958 meeting protocols between 1995 and 2017, which contains 65,393 speeches in total.

Other than the national-level policy-making artefacts, there are other datasets such as EurLex and PreLex which are useful for studying EU politics [Borrett and Laurer (2020); König et al. (2006)]. EurLex provides access to European Union law and other public documents. It has summaries of EU legislation and thematic categorization of legislation. PreLex can be used to monitor the inter-institutional decision-making process — legislative (for example, directives, regulations or decisions), non-legislative (for example, working papers, communications or reports) and budgetary proposals. It tracks the progress of legislative proposals and other policy documents submitted by the Commission to the other EU institutions, covering various events

[13] https://budget.gov.au/2019-20/content/myefo/index.htm

involved during all stages of the decision-making process. Glavaš et al. (2017b) proposed an unsupervised cross-lingual political text scoring approach based on TF-IDF weighed average word embeddings to represent documents and a graph constructed using pair-wise document similarity. Given pivot texts for both left and right ideologies, a label propagation approach is used to predict the ideological position of other documents. They used a corpus of European Parliament speeches to study their approach, and demonstrated that it performs better than the commonly used Wordfish model.

2.2.2.1 Policy agendas

Policy agenda research deals with quantifying issues related to policymakers' decision-making [Karan et al. (2016)]. One of the major efforts in political science is the analysis of policy agendas [Kingdon (1995)]. It includes analysing text from a wide range of political actors, unlike the CMP (see Section 2.2.1.1) which is restricted to manifestos only. Policy agendas research focuses on studying manifestos, media, parliamentary and legislative text, Prime Minister and executive speeches, judiciary documents, budget related text, and data from public opinion and special interest groups. This thread of research have been primary influenced by the Policy Agendas Project (PAP), which tracks policy agenda changes over time [John (2006)]. To this end, PAP developed a codebook for the United States context comprising 19 major topic and 225 subtopic codes, to provide comparable measures of policy changes including: congressional roll call vote, party platforms, public law passed, and news data. Building on the success of PAP, the Comparative Agendas Project (CAP) [Bevan (2019)] extended the PAP codebook, to study policy changes comparatively, across both time and countries. The CAP codebook[14] consists of 21 major topics (given in Table 2.6) and more than 200 subtopics. For example, *unemployment rate*, *interest rates* and *price control* are subtopics under the major topic of *Macroeconomics*. The codebook is used for annotating and studying political texts from over 18 democracies. Consequently, CAP-coded data have been the primary source for many policy agenda studies [Baumgartner et al. (2006)]. CAP aligns multiple sources that are related to policy agenda analysis, compared with CMP which focuses on manifestos only. The choice of sources, which are annotated varies across countries, due to differences in political systems, as does the label schema. For example, the German codebook contains policy agendas related to *German reunification* and *European Union matters*. On the contrary, the label schema is uniform across all countries in the CMP.

[14] https://www.comparativeagendas.net/pages/master-codebook

Code	Topic
1	Macroeconomics
2	Civil Rights
3	Health
4	Agriculture
5	Labor
6	Education
7	Environment
8	Energy
9	Immigration
10	Transportation
11	Law and Crime
12	Social Welfare
13	Housing
14	Domestic Commerce
15	Defense
16	Technology
17	Foreign Trade
18	International Affairs
19	Government Operations
20	Public Lands
21	Culture

TABLE 2.6: Comparative Agendas Project — Major topics

Karan et al. (2016) developed a topic classification model using a dataset based on the Croatian Policy Agendas project. They focused on approaches to handle the hierarchical structure of the topic scheme — one vs one, one vs all, and hierarchical classification taking the confidence of classifier into account to get the final predictions. They showed that the hierarchical model performs better than other approaches. Širinić (2019) examined the agenda setting of weekly government meetings in Croatia. They did an empirical study of issue attention, and number of issues focused on by the government. They found that governments selectively focus on key policy areas such as defence or the economy, and also on a small number of problems. Brandt (2019) target the task of restoration policy agenda classification, with 14 custom categories, across policy documents. They pose the problem as a neural information retrieval task — to classify policy documents based on cosine similarity between paragraphs and query embeddings. They study policy documents in Malawi, Kenya, and Rwanda, and provide analyses of forest and landscape restoration policies.

The United Nations Development Programme (UNDP) helps countries implement the United Nations Sustainable Development Goals, an agenda for tackling major societal issues such as poverty, hunger, and environmental degradation by the year 2030. UNDP provides a review of

national development plans to assess alignment of national targets with one or more of the 169 targets of the 17 Sustainable Development Goals. Galsurkar et al. (2018) targeted agenda and goal analyses of national development plans. This was posed as an information retrieval task again — paragraphs in the documents are embedded to a latent space (paragraph embedding) and matched with the embeddings of Sustainable Development Goals. The proposed system is intended to speed-up the manual review of long policy documents. They also show through their pilot project for Papua New Guinea, that the national plan assessment time can be reduced from an estimated 3-4 weeks to 3 days.

2.2.2.2 Program to policy alignment

Political parties are central actors in democracy, which mediate between societal demands and government policies by formulating policy proposals before elections. Among the various rhetorical functions fulfilled by an election campaign [Eder et al. (2017)], perhaps the most important is the contract they represent between parties and voters in terms of promises and prioritisation of political issues [Royed et al. (2019)]. Fulfilment of election pledges is critical to the concept of *representative democracy*. It is also a keen area of focus by political scientists, in linking election promises to government programs and policies, also known as program-to-policy linkage [Royed (1996); Thomson (2001); Schermann and Ennser-Jedenastik (2014)]. Program-to-policy linkage or alignment refers to the congruence between government policies implemented and the promises made by political parties to voters during election campaigns, often set out in their election programs or manifestos. Political scientists have found that if parties channel societal demands into public policies effectively, then the level of congruence between election manifestos and government policies should be high [Thomson et al. (2010)]. Various factors such as single-party parliamentary systems (e.g., US) vs. coalition (e.g., Germany), ideological position, issue salience, and institutional constraints (e.g., European Union membership) have been shown to influence the pledge fulfilment [Thomson et al. (2017); Pétry and Duval (2018); Klingemann et al. (1994)]. For this alignment, a preliminary step is to collect and analyze election promises. Towards that, the Comparative Party Pledges Project provides the largest collection of election pledges [Naurin et al. (2019)]. All these efforts have been carried out manually, there are no publicly available curated resource for developing automatic techniques. This forms the motivation for one of our research goals in this thesis, which deals with automatically tracking progress made towards election promises (Section 3.4).

FIGURE 2.3: Example petition from UK Government Petitions website.

2.2.3 Online petitions

A petition is a formal request for change or action to any authority, co-signed by a group of supporters. In general, the target authority for a petition can be political (e.g., We The People in US) or non-political (e.g., Change.org). The target of political scientists is generally platforms run by governments, so that people are able to engage with elected representatives directly, which is commonly referred to as *advocacy democracy* [Dalton et al. (2003)]. The European Parliament's definition of petitions is [Böhle and Riehm (2013)]:

> *as all complaints, requests for an opinion, demands for action, reactions to Parliament resolutions or decisions by other Community institutions or bodies forwarded to it by individuals and associations.*

Typical steps involved in the working of electronic petitions are:

Creating petitions: Petitions are accepted either via e–mail or using a web interface. During this initial submission phase, the petition text can be supplemented with additional background information — such as the author's name, email address and postcode [Lindner and Riehm (2011)]. An example petition from UK Government Petitions platform,[15] is given in Figure 2.3.

Gather support: After the initial submission of petitions, they may be manually checked for compliance with standards, and the authors asked for additional details when incomplete. Petitions can also be rejected if they do not meet the standards expected by platforms. For example,

[15] https://petition.parliament.uk/

The Government recognises the strength of feeling around the use and misuse of fireworks and has listened to the concerns raised in parliamentary debate and wider discussion. We receive representations from a wide range of stakeholders, including members of the public, organisations and charities, with wide-ranging views on what the issues are and what action they would like to see.

Following the Westminster Hall debate on 26 November 2018 regarding fireworks, the Minister with responsibility for fireworks policy and legislation in the Department for Business, Energy and Industrial Strategy, Kelly Tolhurst MP, asked the Office for Product Safety and Standards (OPSS) to develop a fact-based evidence base on the key issues that had been raised. This includes looking for data around noise and disturbance, anti-social behaviour, non-compliance, environmental impact, and the impact on humans and animals. As part of this work we are considering the findings of the Scottish Government consultation on fireworks, which was published on 4th October. We will also consider the House of Commons Petitions Committee inquiry on fireworks once that has reported.

The aim of the evidence base is to build a full picture of the data around fireworks in order for government to identify whether there is a problem, and if so, what action - if any - is appropriate. This work will also help us identify trends across fireworks seasons and determine whether, there has (for example), been an increase in fireworks being set off or an increase in firework related injuries.

Department for Business, Energy and Industrial Strategy

FIGURE 2.4: Snippet from UK Government's response to the petition given in Figure 2.3.

in the case of UK Government Petitions, *petitions must call for a specific action from the UK Government or the House of Commons*, and there are many more conditions.[16] Once published, subsequent steps depend on the support it gets from other citizens. This is often realized with co–signing of petitions. Other possibilities, though not very common, are discussion forums which allow public to debate on the issues raised by a petition (e.g., Bundestag e-petitions[17]) [Lindner and Riehm (2011)].

Government action: Once a petition obtains the required amount of support, then the government provides its response action. In the case of UK petitions, petitioners are guaranteed an official response at 10k signatures, and the guarantee of parliamentary debate on the topic at 100k signatures; in the case of US petitions (We *the* People[18]), they are guaranteed a response from the White House at 100k signatures. An example snippet of a response from the UK Government to the petition in Figure 2.3 is provided in Figure 2.4.

E-petition systems vary across countries including: Europe [Böhle and Riehm (2013)], UK and the US. Research has shown the impact of online petitions on the political system [Lindner and Riehm (2011); Hansard (2016)], in terms of their increasing popularity, and the integration of

[16] https://petition.parliament.uk/help

[17] https://epetitionen.bundestag.de/petitionen/_2019/_10/_24/Petition_100651/forum/Beitrag_638864.nc.html

[18] https://petitions.whitehouse.gov/

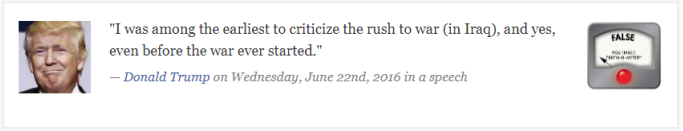
the public into the political system. Though it is observed that petitions are not an effective means for holding governments accountable [Schmitter and Trechsel (2004)], they are initiated bottom–up by citizens with low access barriers, which compels the attention of the addressee [Böhle and Riehm (2013)]. Other than these government petition platforms, there are also general petition platforms such as Change.org, Avaaz.org and ThePetitionSite, which are written and signed by people across countries (not just one country's citizens) and the usual targets are government agencies and business organizations [Elnoshokaty et al. (2016)].

Modeling the factors that influence petition popularity — measured by the number of signatures a petition gets — can provide valuable insights to policy makers as well as those authoring petitions [Proskurnia et al. (2017)]. Previous work on modeling petition popularity has focused on predicting popularity growth over time based on an initial popularity trajectory [Hale et al. (2013); Yasseri et al. (2017); Proskurnia et al. (2017)], e.g. given the number of signatures a petition gets in the first x hours, prediction of the total number of signatures at the end of its lifetime. Hale et al. (2013) targeted UK government petitions, and found that most successful petitions grow quickly, and the number of signatures a petition receives on its first day is an indicator of the total number of signatures it receives during its lifetime. Yasseri et al. (2017) examined early growth within the first day using UK government petitions data at an hourly resolution. They found that the signature distribution was highly skewed, and the intrinsic time scale of the platform is very short, where growth of signatures exhibits rapid dynamics (rate of growth or decay). They also suggested shorter deadlines for petitions, similar to Germany (three/six weeks) or US (one month), and argued that it might produce similar outcomes without the clutter of old petitions. Proskurnia et al. (2017) proposed a time-series regression model, by extending the Hawkes process based approach [Kobayashi and Lambiotte (2016)], that incorporates seasonality, aging effects, self-excitation, and external effects. They test the method using ThePetitionSite data, and find that the success of petitions is influenced by several factors, including anonymity of the authors, promotion through social media channels, and whether a petition appears on the front page of the petitions platform.

The Petitions Committee uses Twitter to engage the wider public in the parliamentary debates of petitions (for those petitions that reach over 100,000 signatures). Asher et al. (2017) study Twitter conversations related to UK petitions, in terms of people's engagement with e-petitions beyond signing them, how people perceive the e-petition procedure, and who takes part in these conversations. The authors found that the discussions are often emotional and polarized. Emotion was analyzed using a lexicon based sentiment analysis technique [Nielsen (2011)], and

polarization by training a Naïve Bayes classifier using a manually annotated tweets dataset. [Hagen et al. (2015)] studied US petitions (We The People), and extracted three features — informativeness (unique number of words in each petition), named entities, and topics (using a topic model) — and show via a regression model that the features (informativeness, named location, and several topics) have a significant correlation with the signature counts. Elnoshokaty et al. (2016) analyse Change.org petitions and perform correlation analysis of popularity with the petition's category, target goal set, and the distribution of words in General Inquirer categories [Stone et al. (1962)]. Huang et al. (2015) analyze 'power' users, based on the number of petitions they sign, on the Change.org petition platform, and show their influence on other petition signers. They also find out that the top 5% of Change.org users contribute to more than 50% of all signatures. It should be noted that the work of Huang et al. (2015); Elnoshokaty et al. (2016) and Proskurnia et al. (2017) uses general petitions (which includes political petitions), but not those addressed to governments specifically; however the techniques they developed are relevant to the popularity analysis task in a political setting.

Hale et al. (2018) studied how changes to the design of the UK government petition platform can impact public participation. The authors test the hypothesis that social information, in the form of the *trending feature* given by a petition's rank on the homepage, shapes the distribution of political mobilization. They found that the *trending feature* did not increase the volume of signatures, but changed its distribution by allowing popular petitions to gain more signatures at the cost of other petitions with fewer signatures. In addition, the UK Parliament's Petitions committee geo-locates petitions according to the location of the signatory. This data is available for each petition — counts allocated to a country (if outside the UK) or to a constituency. Clark et al. (2017) used the UK petitions data to characterize constituencies, based on the number of signatories in each constituency. This is based on the idea that political sentiment in each constituency can be captured using the volume of signatories to e-petitions [Wright (2015)]. Using Gaussian finite mixture model they identified four clusters and also show that they are coherent in terms of the political discourse. For example, a conservative class of petitions was concentrated in the urban centres. They also showed how the output clusters map well onto the 2017 UK general electoral results. Similarly Vidgen and Yasseri (2020) study UK petitions submitted during 2015-2017 (surrounding the UK EU-membership referendum). The authors investigate the interaction between topics (using a topic model), temporal dynamics, and geographic features. They show that the popularity of some issues are stable over time (e.g., healthcare), while others are influenced by external events, such as the referendum in June 2016 (e.g., international



Trump still wrong on his claim that opposed Iraq War ahead of the invasion

FIGURE 2.5: One of the Trump's claims verified by PolitiFact.

affairs, EU). They also show that some issues receive support from regions across the country (e.g., law & order, work & pay) but others are far more local (e.g., animals & nature, driving). Further they cluster constituencies based on the issues signed by constituents. They also validate the clustering output by aligning eight of the ten petition issues with the top ten issues reported in Ipsos MORI survey, which includes monthly political surveys to study political trends.

2.2.4 News data

News sources contain various political events, views and opinions on different issues from different actors in a democracy. This helps in understanding facts and opinions related to national political issues. News also contains international events, which enables understanding of international relations. For instance, the GDELT (Global Data on Events, Location, and Tone) project monitors streams of broadcast, print, and web news across countries in over 100 languages, and identifies information such as entities (people, locations, organizations), themes, emotions, and events, which can enable applications such as monitoring civil unrest triggered by government policies [Sharma et al. (2017)]. Halterman (2019) developed a neural network based approach for geolocating political events in news, which can be used to analyze news events both at national and sub-national levels.

News media covers party pledges from major political parties, though some policy areas are better represented than others [Kostadinova (2017)]. It also tracks policy updates related to election pledges. Tan et al. (2018) studied media highlights of US Presidential debates. They quantitatively explore what factors influence media selection, using a dataset which contains Presidential debate transcripts and post-debate coverage. They find that choices of wording — informativeness, emotion, contrast, uncertainty, rhetorical technique and so on — have significant influence. They also show that human annotators do not achieve a high accuracy (only 60%), indicating that media choices are not entirely obvious. Duval (2019) show that news

media alerts citizens when a pledge is broken. Secondly, journalists from organizations such as PolitiFact manually check the veracity of claims made by politicians during election campaigns and even after government is formed. For example, Figure 2.5 shows one of Trump's claims verified by PolitiFact, where the journalists rate Trump's Iraq war-related claim, made before the 2016 elections, to be false. In addition, they also keep track of progress made towards promises by the party which forms government after elections. An example from PolitiFact which tracks Trump's 2016 election promises is given in Figure 2.6. This is related to *program-to-policy alignment* (Section 2.2.2.2), and here is done by journalists in a news organization and the data is available for analysis. These systems are critical because campaign communication provides the medium to connect political parties with voters, and also influence voters' decision making [Fernandez-Vazquez (2014b)]. The analysis has mostly been done manually. As an exception to this, Barrón-Cedeño et al. (2018) targeted the task of automatically verifying the truthfulness of claims made during 2016 US presidential campaign (aka fact-checking). The related Conference and Labs of the Evaluation Forum (CLEF) 2018 task has two components: (a) finding and ranking evidence sources according to their usefulness towards fact-checking a target claim (evidences can be news articles and other web-pages); and (b) using the evidences to predict the factuality of claim. Popat et al. (2018a) handled the task by obtaining evidence via searching the web (especially for news) and social media sites, and Popat et al. (2018b) used the top search results from Bing. Towards the fact-checking task, a preliminary step is to find check-worthy claims from campaign text, e.g., presidential debates [Hassan et al. (2015); Atanasova et al. (2018)]. Atanasova et al. (2018) is again a CLEF 2018 task based on 2016 US presidential campaign text for claim check-worthiness estimation. Hassan et al. (2015) created a dataset based on presidential debates, using lexical, sentence length, POS tags, entity types, and sentiment features, with supervised learning approaches. Audit of both claims and the functioning of government (post election) is very much essential to improve trust of voters in democracy. Overall, such audit systems enable the accountability of political actors in the electoral system both during and after the elections.

Lerman et al. (2008) studied how news articles shape public perception of political candidates. They evaluate the predicted shifts in public opinion to what is reflected in prediction markets (e.g., TradeSports, Iowa Electronic Markets[19]). They used a Logistic Regression model with hand-crafted features such as bag-of-words, news focus (lexical overlap between debate text and subsequent days news), dependencies labeled with grammatical function labels, and named

[19] www.geekwire.com/2016/hillary-clinton-slumps-donald-trump-jumps-political-prediction

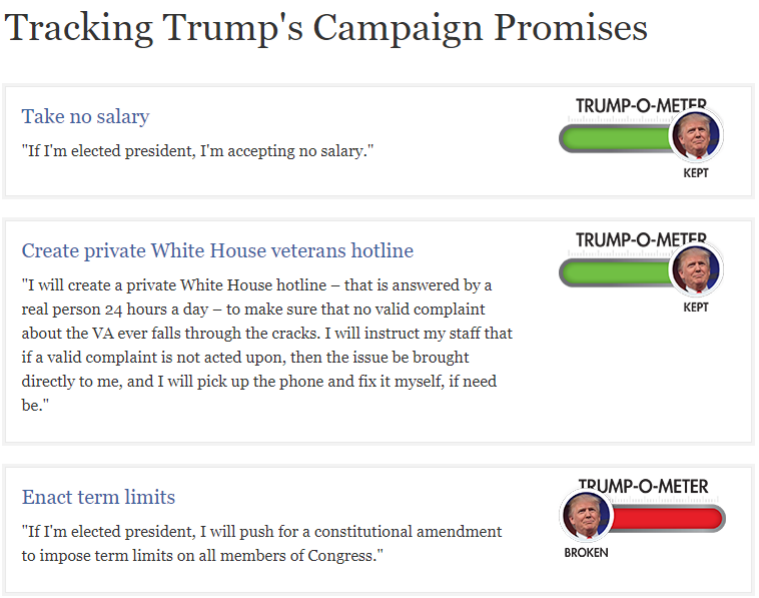


FIGURE 2.6: Sample set of Trump's election promises tracked by PolitiFact.

entities. Many political news sentiment tasks handle resource-rich languages such as English, whereas Bakken et al. (2016) targeted lower-resource languages such as Norwegian. They used supervised learning algorithms using bag-of-words and negation features.

Lastly, in online news, the same event can be presented from very different perspectives, depending on the source. Subtle differences in opinion expressed by journalists on different issues [Bakken et al. (2016)], can help to uncover media bias. For example, AllSides[20] assesses media bias or political leanings of hundreds of media sources. They are based on blind surveys of people across the political spectrum, and a variety of other survey information. Zhou et al. (2011) propose a semi-supervised graph based approach to classify news articles and users as liberal or conservative, modeling homophily in the data. They use news opinion articles and readers' subjective reactions (votes) to those articles via news aggregator services. A graph is constructed with articles and users as nodes, and links representing the votes by users for particular articles. Given a few labeled articles and users, they use a graph based approach to predict labels of the remaining users and articles. Gangula et al. (2019) focused on the task of detecting news article bias towards a political party. Their dataset contains headlines and the content of articles written in the Telugu language, annotated for the political party towards which it is biased, or *none* if it is not biased. More often bias is subtle and related to the framing of issues, which is an important part of determining how information will be interpreted [Card et al. (2015)]. Card et al. (2015) investigated the relationship between framing and news, and released the Media Frames Corpus,

[20] https://www.allsides.com/media-bias/media-bias-ratings

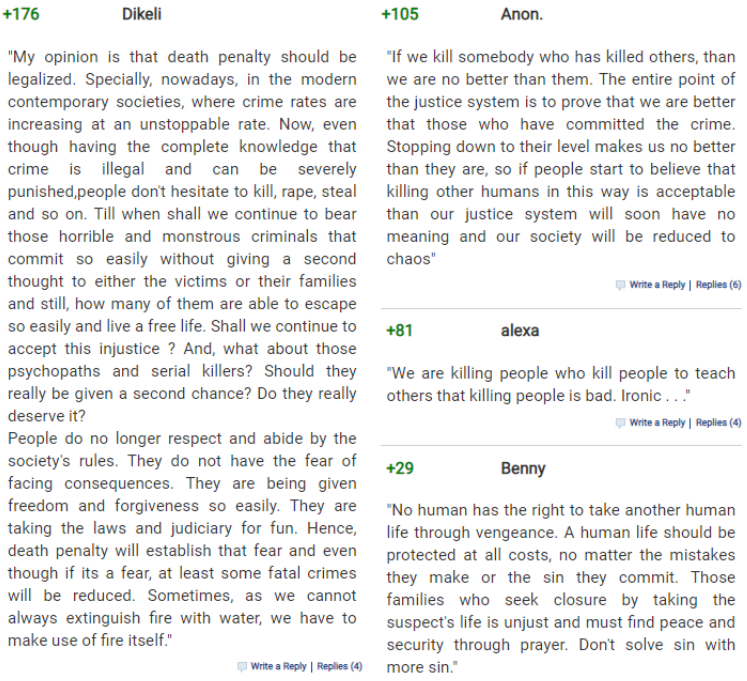


FIGURE 2.7: Snapshot taken from procon.org. Issue discussed is *Should the Death Penalty Be Allowed?*. Pro discussions are given on the left-side, and Con are given on the right-side.

FIGURE 2.8: Twitter campaign for a petition written to UK Government.

which contains news articles annotated with framing dimensions such as "fairness and equality". Fan et al. (2019) studied bias at both the document- and sentence-levels using news articles from different news organizations (e.g, Fox News). At the document-level, manual annotations are obtained to estimate the overall sentiment towards the main event, based on which the articles from different news organizations are ranked on the ideological spectrum. At the sentence-level the task and annotations are more nuanced, involving identification of the bias type (*lexical* or *informational*), the target of the bias, bias polarity, whether it is a direct or indirect bias towards main target, and whether the bias is part of a quote. *Lexical bias* is captured based on the use of language, whereas *informational bias* is based on the factual content conveyed in a way to influence the reader's opinion.

2.2.5 Social media discussions

Social media such as blogs, discussion forums, Twitter, and Facebook have garnered public attention for their use in different socio-political events [Diakopoulos and Shamma (2010)]. In the recent decade, there has been increased use of digital media for communication and coordination of political events, including election campaigns; policy discussions and feedback; politicians sharing their views and opinion on various policy issues; advocating for change in policies (see Figure 2.8); and discussing and sharing perspectives (see Figure 2.7). The conversations can be between party representatives and voters, between voters, or between politicians. They include political discourse in various scenarios, including a case where discussions are based on a news item [Lee (2018)]. In effect, social media can be seen as a medium to connect various actors in a democracy, not just faster but also in a decentralized fashion.

Researchers have used social-media data for analyzing candidates' popularity during election cycles [Karami and Elkouri (2019)]. The popularity of a politician is seen as a critical factor for the politician or their party to win elections [Gaurav et al. (2013)]. Gaurav et al. (2013) show that the number of times the name of a candidate is mentioned in tweets prior to elections correlates with their success in elections. Beyond popularity, Twitter sentiment classification is a popular task. Bakliwal et al. (2013) studied sentiment of tweets towards a particular political party or party leader, during the run-up to the 2011 Irish General Elections. They used supervised learning with subjectivity and tweet-based features. Ringsquandl and Petkovic (2013) targeted aspect-based sentiment analysis with tweets related to the 2012 US Presidential election. They use a semi-supervised approach to build a domain-specific sentiment lexicon by expanding a general lexicon [Wilson et al. (2005)], then used a lexicon-based approach for both sentiment classification and opinion summarization. Political tweet sentiment has been investigated to predict poll outcomes [Bermingham and Smeaton (2011); O'Connor et al. (2010); Tumasjan et al. (2010)]. Bermingham and Smeaton (2011) use supervised learning approaches with manually labeled data, O'Connor et al. (2010) use a lexicon based technique [Wilson et al. (2005)] and moving average to track opinion over time. Tumasjan et al. (2010) analyzed political sentiment also using a lexicon based approach (Linguistic Inquiry and Word Count: [Pennebaker et al. (2015)]). Political tweet data has also been utilized to predict users' voting intentions for political parties [Lampos et al. (2013)], modeling both text and the users, framed as a bilinear model. The authors also evaluated a multi-task extension to model related output variables, in their case inferring voting intentions for competing political parties.

Apart from candidate's mention-based popularity and sentiment, an important task is stance classification — determining whether an individual is in favor of, against, or neutral towards a target concept, given the content they have generated. Johnson and Goldwasser (2016) analyzed how issues are framed in individual tweets and temporally to collectively infer the most likely stance. For collective inference, they use probabilistic soft logic [Bach et al. (2017)] to capture both local and global dependencies using keyword based approaches to classify issues [O'Connor et al. (2010)] and sentiment [Wilson et al. (2005)] (unary predicates), alongside inter-dependency between sentiment and stance, and agreement (or disagreement) between similar politicians' (based on tweet content) stance. Joseph et al. (2017) deal with Twitter stance classification. Since the amount of context can affect the quality of annotation, to characterize and reduce these biases they propose a probabilistic model which infers ground truth labels as well as learns a classifier. Hasan and Ng (2013) target debate stance classification — which is to classify the stance expressed by an author from a post written, in the context of a two sided debate in an online debate forum (similar to Figure 2.7). The authors propose an approach using text features — lexical, syntactic, sentiment and argument features [Anand et al. (2011)] — as well as user-interaction constraints and ideology constraints. Adamic and Glance (2005) studied linking patterns and discussion topics of political bloggers — specifically, how often liberal and conservative (community) blogs referred to one another and the overlap in the topics discussed — both within the liberal and conservative communities, and also across communities. The authors found a divided blogosphere with very few cross-links between communities, but not similar homogeneity in the news and topics discussed.

Johnson and Goldwasser (2018) model how politicians frame issues on Twitter to predict the moral foundations used to express their stance on different issues. Moral foundations include care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/degradation. Politicians express their moral and social positions on issues with the goal of maximizing the response to their messages. They studied the use of moral foundations via word-list and domain-specific unigrams, to jointly model policy frames and moral foundations. Wei et al. (2013) analyzed tweets to study the influence of mainstream media towards shaping public opinion during the 2010 UK General Election. They quantify media bias using sentiment analysis, where they show that the attitude correlates well with political bias in the UK media. They also found that tweets from individual journalists tend to have a superior reach compared to tweets from mainstream media.

Proskurnia et al. (2017) showed the effect of Twitter on petitions' success using Granger causality test between tweets related to the petition and its signatures (similar to the example shown in Figure 2.8). They showed the predictive accuracy improves when using tweet data, especially to reduce the amount of data required for accurate predictions. Proskurnia et al. (2016) showed that environmental petitions' signatures and Twitter shares have a positive correlation. Aragón et al. (2018) also showed that there is a positive correlation between shares on social media platforms, and the number of signatures for a petition, using a dataset constructed from the Avaaz.org petition platform.

Twitter is used for other political activities besides campaigning for petitions. Hemphill and Roback (2014) study how constituents use Twitter to lobby their elected representatives about issues that matter to them. The authors first search Twitter for hashtags that are related to popular political issues, and then manually annotate them for lobbying strategies given by speech acts [Austin (1962); Searle (1976)] — directive, commissive, representative, expressive, and questions. From investigating the data, they show that citizens don't use Twitter to merely shout out their opinions on issues, but utilize various lobbying strategies to impact political outcomes. Shapiro et al. (2018) analyze how members of the US Congress use a range of speech acts to engage in discussions on the Twitter. Specifically, they study whether the speech act distribution changes over time; and how speech act usage differs between men and women, between chambers, and between parties. All these approaches analyze speech acts in tweets, but are not specifically tied to the policy proposal or policy-making stages. In this thesis, we model speech acts of political campaign text with a particular focus on pledges (see Section 3.2), which is crucial for verifying their fulfilment.

2.3 NLP challenges

A primary challenge when applying NLP to political analysis is to first create the task definition (for unexplored challenges), and/or annotate datasets when there are no publicly available benchmarks (as commonly occurs). Secondly, there are various challenges related to analyzing the different sources of text discussed previously in this chapter. They arise for multiple reasons: annotations require domain knowledge; properties of text vary across sources, particularly so across countries, which makes it difficult for a classifier trained on one source to perform well on other sources; political text often use visual media such as graphs and tables to convey information; it is difficult to incorporate the rich domain knowledge available from decades

of political science research and develop interpretable models to further data understanding for political scientists; and finally political speech can be multi-modal by nature, including text, speech and video signals.

In order to automate NLP tasks, a core challenge is in coming up with a task definition, which requires collaboration between NLP and political scientists, and a good understanding of the political science literature. After defining the task, it is necessary to obtain annotated data for building NLP models. Since it is expensive to get large-scale annotations for text in niche domains such as politics, it is necessary to develop algorithms that can cope with little or no supervision, and leverage the often abundant unlabeled text. Political scientists have proposed unsupervised text scaling approaches such as Wordfish [Slapin and Proksch (2008)] which is a popular technique for obtaining ideology scores of documents based on relative frequency of words. Johnson and Goldwasser (2016) use weakly supervised learning for classifying politicians' stance on different policy issues using Twitter discourse; and in Charalampakis et al. (2016) use semi-supervised learning to detect irony in Greek political tweets. In Karan et al. (2016) model topic agendas in media text, using a semi-supervised approach involving a topic classification model. They first estimate the confidence of classifier decisions for each individual document, and then forward the documents with low-confident decisions to a human coder. Zhou et al. (2011) propose semi-supervised label propagation techniques to classify news articles and users as liberal or conservative (exploiting homophily).

After obtaining labeled data for some sources, in order to compensate for the lack of labeled data in other sources we can employ *transfer learning*. But existing work has observed that this is non-trivial [Gentzkow and Shapiro (2010); Yan et al. (2017)]. A major reason for this is the difference in vocabulary used to refer to the same concepts. For example, in US debates over privatizing social security, liberals (Democrats) typically used the phrase *private accounts* whereas conservatives (Republicans) used *personal accounts*, but a news text on this topic could use different phrases again [Gentzkow and Shapiro (2010)]. Yan et al. (2017) study domain adaptation for the ideology prediction task using congressional records, political web-pages and wiki text. They observed that the cross-domain learning performance was poor even though the models perform well in the in-domain setting. Even with a single corpus, it can be hard to build robust models due to changes in vocabulary over time. Desai et al. (2019) proposed adaptive ensembling for handling diachronic corpora, where the source domain is labeled and the target is unlabeled. They use a student-teacher framework, which has a training loss and consensus loss. The teacher network's parameters form an ensemble of the student network's parameters

over the training epochs. Rather than using weights for parameter sets, Tarvainen and Valpola (2017) introduce learnable weights for each parameter in the student network.

Other than vocabulary mismatch, domain adaptation can also suffer from *confound shifts*, i.e., there exists a confounding variable that influences both the input text and the target variable. For example, the *healthcare* topic might be more affiliated to one political party at time t and to a different party at time t+1. Landeiro and Culotta (2016) propose a framework to control for confound variables during training, by introducing separate regularization term for confound variables in learning. Landeiro and Culotta (2018) extend the framework to handle multiple confound variables. They evaluate on political party affiliation prediction and other non-political tasks. Their political tasks include: (a) predicting the party affiliation of a member of the Canadian parliament based on the their floor speeches, where the confound is whether the speaker belongs to the governing or opposition party; (b) predicting the party affiliation of a member of US Congress based on their Twitter content, where the confound is the topic of discussion. Previous work has assumed that the confound variables are known, which is not practical in all applications. Landeiro et al. (2019) focus on discovering confounds using a topic-model based approach, and handling confound variables using an adversarial learning technique which fits a model that is invariant to the latent confounds.

Apart from analysis targeting political text from one country, covering multiple countries is a key requirement for comparative political text analysis. In many cases, data from different countries is written in different languages. Hence, handling multi-lingual data is a very important challenge for achieving comparative political text analysis. Secondly, different political studies have taken place around the world, so it will be potentially necessary to leverage expert surveys or other annotations in one language to build machine learning models for other (e.g., low-resource) languages. Existing work handles this using a range of solutions, from machine translation to obtain text in a common language [Windsor et al. (2019)] to using multi-lingual modeling approaches [Lucas et al. (2015)]. Bilbao-Jayo and Almeida (2018) use Word2Vec trained in different languages for manifesto text political theme classification in various languages. Glavaš et al. (2017a) handle this challenge using multi-lingual embeddings projected on to the same space.

In addition to handling label sparsity, in general political analysis requires handling multiple sources of data. For instance, DIME (Database on Ideology, Money in Politics, and Elections) [Bonica (2019)] is a resource for studying campaign finance and donor ideology in American

politics. It contains over 130 million political contributions made by individuals, interest groups and organizations to several elections spanning the period of 1979-2014. Curating and analyzing such a database requires handling data from multiple sources, especially with different characteristics such as length and writing style.

There are also various important (and relatively less explored) resources, such as bills and budget documents (e.g., Australian budget 2019-20 statements), present data not just using text, but also using lists, tables and figures. Parsing and using non-textual information jointly with text can be helpful, as in the case of scientific literature [Clark and Divvala (2015)]. This can help to understand fine-grained details from the document, and trends based on budget details [Laurie et al. (2008)]. More importantly, any deeper analysis with such documents requires representing and reasoning over numbers. Wallace et al. (2019b) study how various token embedding methods can capture numeracy using the DROP dataset [Dua et al. (2019)]. They observed that ELMo [Peters et al. (2018)] captured numeracy better than other pre-trained models.

Moreover, political speech is highly multi-modal in nature, and can benefit much from other modalities in the data. In particular, voice is seen as a strong indicator of the affective states of the speaker [Campbell (2007)]. Lippi and Torroni (2016) showed that the performance of claim detection for argument mining can be improved by employing both text and speech features together, which they demonstrated on the 2015 UK Prime Ministerial election debate data. Gillick and Bamman (2018) focused on modeling applause in campaign speeches, and showed that lexical information is most informative, but audio is also required to capture acoustic features. Acoustic features including pitch, energy, and a broader pitch range are all predictive of applause. While past work has focused only on textual indicators of applause, this work showed that audio signals or how a message is delivered is equally important.

Lastly, an important set of challenges includes incorporating domain knowledge into models, and building interpretable models which can be verified, and can answer the questions of domain experts, and thereby enable expansion of existing knowledge. Johnson et al. (2017) and Johnson and Goldwasser (2018) leverage party bias towards policy issues to study stance and moral foundations. Kornilova et al. (2018) leverage partisanship (or party-level homophily) through sponsorship data to model legislation voting decisions in the US Congress. In terms of interpretability, Tan et al. (2018) study what factors influence media selection of highlights from Presidential debates. Apart from intrinsic text feature based analysis, they show through crowdsourcing that media choices are not entirely obvious. This opens a different dimension of

challenge which involves both the predictability as well as the interpretability of models [Miller (2019)].

2.4 Summary

In this work we focus on applications related to analyzing policy proposal (analysis of manifesto, speech transcripts and media-releases); tracking policy implementation and progress towards election promises; and analyzing petitions to understand voter advocacy. In the context of this thesis, we address the following:

- we handle multi-lingual text for fine-grained policy classification and ideology prediction of manifestos (Section 3.1).
- We handle semi-supervised scenarios by using multi-task learning for manifesto analysis, and use a student-teacher semi-supervised learning framework for campaign text speech act classification and promise specificity prediction (Sections 3.1, 3.2 and 3.3).
- For the other two tasks — tracking progress of promises, and petition popularity analysis — we use standard supervised learning approaches, and the major contribution lies in automating these tasks (Sections 3.4 and 3.5).

Other than manifesto and petition analysis, all the tasks we target are novel to the NLP community. For these tasks, we have curated data from multiple sources, and in several cases annotated the data manually by crowd-sourcing. We also incorporate domain knowledge, in the form of coalition and temporal dependencies, following Greene (2016) (Sections 3.1 and 3.3), and study factors that influence petition popularity by interpreting the predictive model (Section 3.5).

Chapter 3

Research Contribution

The research contributions of this thesis span methods for the analysis of different sources of political text. Our goal is to automatically analyze text generated by political parties, their representatives, and voters, in the context of improving political accountability and voter advocacy. The individual research goals correspond to key political science questions, which involve extensive manual coding to answer. Note that, in most cases, political scientists care about the downstream political analysis. For example, if a party makes more promises on *welfare issues* compared to *economy*, then that party is more *left-leaning*. This analysis requires classifying the policy theme of sentences and their rhetorical functions (to know if it is a promise or not), and these steps are usually done manually. The goal of this thesis is to automate these text predictions, and at the same time, wherever applicable, perform downstream analysis to evaluate the political scientific conclusions.

The thesis contributions can be logically split into three chunks. The first research contribution deals with CAMPAIGN ANALYSIS — how to automatically analyze election campaign text, which can help voters to keep track of parties' policy proposals. Towards this, we analyze policy issues discussed in campaign materials, and the context of each utterance — whether it is a belief statement or a reference to a past action or a promise, which can help to uncover several party position indicators studied by political scientists (Section 2.2.1.1). As discussed in Section 2.2.1, campaign text is available through different medium including election manifestos, speeches, and media-releases. Manifestos are a core artefact used by political parties to summarize their goals before elections, while speeches and media-releases contain more dynamically evolving policy goals through the election cycle. We use manifestos to study policy theme and

Text source	Task	Theme
Manifestos	Policy theme analysis	Campaign analysis
Media-releases and speeches	Speech acts classification	
Manifestos	Pledge specificity prediction	Promise tracker
News and media-releases	Promise tracker	
Petitions	Popularity prediction	Voter advocacy

TABLE 3.1: Research contributions summary

party ideology as they are a self-contained resource, and also popularly used by political scientists for this purpose. We use media-releases and speech transcripts to study rhetorical functions in political discourse, as they are dynamic in nature where politicians use several types of utterances to convey both the political function as well as the general persuasive function [Gautrais et al. (2017)], in an attempt to win elections.

The second goal of this thesis is PROMISE TRACKING — how can we automatically identify verifiable promises, and build a system which can keep track of progress made towards election promises after a government is formed. This connects both policy proposal and policy making stages involved in a policy process (Section 2.1). This can be seen as a way to automate program-to-policy alignment (Section 2.2.2.2), which can help to improve the accountability of political parties. We use manifestos to identify promises, and analyze them based on specificity, to determine how readily verifiable they are likely to be. Following that, we use news reports and media-releases from relevant sources to keep track of progress made towards those promises.

The third and final thesis part is related to VOTER ADVOCACY — how to automatically predict the popularity of petitions. Among several avenues available for engaging voters in the policy making process, petition platforms run by governments provide a unique channel for voter advocacy (Section 2.2.3), which are initiated bottom-up by citizens with low access barriers. This research goal helps to forecast the popularity of a petition, and thereby determine the breadth of support for the issue, in addition to providing feedback to the petition authors as to the likely impact of the petition in its current form.

The individual research contributions in order to achieve the overall goal is summarized in Table 3.1. Each contribution is associated with a published paper which is included in the thesis, as follows:

- Automating manifesto analysis — fine-grained policy theme classification at the sentence-level and coarse grained ideology prediction at the document-level. The document-level task, which is to quantify the manifesto on the left-right spectrum, is usually seen as a downstream application after obtaining sentence-level labels, and is of more interest to political scientists. The intuition behind the quantification is, if a party discusses more *health funding* than *economy* policies, then it could be left-leaning. We propose deep learning approaches to automate the sentence- and document-level tasks, and rather than modeling the two tasks separately we evaluate the effectiveness of modeling them jointly. We also build a model which not only works for English text, but also for a broad range of Western European languages. We describe our proposed approach and evaluation results in Section 3.1; a preliminary version was published at ALTA 2017, and an extended version at NAACL-HLT 2018.

- Apart from classifying the policy themes of sentences, it is valuable to know the context or rhetorical roles of text. For instance, if a campaign text is about *education* policy (which we can get from the previous task), its function is different if it is an *assertive* (belief) or a *promise*. We handle this by modeling the speech acts of text. This work was published at *SEM 2019, and is described in Section 3.2.

- We now turn to the core of election campaign — promises, which set the expectations of voters, and influence their voting decisions. In terms of influence on voter behaviour it is reasonable to argue that vague promises are of limited utility, while specific ones are more important in terms of accountability. Hence automatically identifying and grading promises in terms of specificity is important for political scientists. Towards this, we propose an ordinal regression model to automate the task of election promise specificity prediction. This work was published at EMNLP-IJCNLP 2019, and is described in Section 3.3.

- After promises have been extracted from manifestos, a core part of democratic accountability is to check whether political parties have kept their election promises. For this, we propose the new task of tracking progress of election promises. We use data curated by journalists and propose a two-stage task to handle this challenge — first pick trackable promises from text associated with an election campaign, then track the progress of each promise given suitable evidence. This work is under submission to Computational Linguistics, and is described in Section 3.4.

- All the previous contributions focused on political parties, but in the final contribution we focus on the primary stakeholder, the *voters*, for and by whom a democracy functions. Among various media, *online petitions* allow citizens to connect directly with their elected representatives. Here we want to predict the popularity of petitions, captured by the number of voters who support a change or request, given the petition text. This can help both the petition owners as well as the platform moderators to improve their campaign strategies, thereby enhancing public engagement. This work was published at ACL 2018, and is described in Section 3.5.

The publications listed above form the core of this thesis. In the following sections, we list those publications, and for each publication we provide (1) the full publication record, and (2) motivation or background for the section. Though the research goals are addressed in an independent fashion, they are all interconnected, and we provide end-to-end analysis in Chapter 4.

3.1 Manifesto analysis [ALTA 2017, NAACL-HLT 2018]

Manifestos are a natural starting point for political research, because they are a canonical resource of party policy proposals, and are easily available through the Comparative Manifesto Project (CMP) [Volkens et al. (2011)], which provides a large digitized collection of manifestos across many countries. The dataset also contains sentence-level annotations for policy themes, coded by domain experts. The sentence-level policy categories are typically used by political scientists to arrive at aggregated measures in order to characterize political parties. One such aggregated score, which is referred to as Left-Right scale (RILE) (discussed in Section 2.2.1.1), quantifies the party position on the *left-right* spectrum. An important aspect about the CMP dataset is that, since not all manifestos are available in a machine readable form, there are more documents with these aggregation scores (including RILE), than ones with sentence-level annotations. Apart from the aggregated RILE metric, which is available for a large collection of manifestos across countries and time, there are also expert surveys which are considered the gold-standard by political scientists. Expert surveys are conducted periodically, by having experts answer questionnaires about parties and aggregating their responses. A major differentiating factor with the expert surveys is that they are contextual, i.e., consider global political shifts, coalitions, and other relevant factors.

For automatically analyzing manifesto text, our research questions are: (a) how can we model the tasks jointly? since sentence-level annotations are used to compute document-level ideology scores, the joint model can capture the inter-dependencies between tasks; (b) can document-level RILE scores be used as supervision for learning sentence-level policy themes classifier under semi-supervised settings?; and (c) though RILE is available for a large collection of manifestos, expert surveys are more sparsely available yet are more trusted among political scientists; can we infer expert scores by suitably incorporating additional contextual information into the model?

We first evaluated a simple hierarchical neural network model with average word embeddings for the joint sentence- and document-level tasks. This work was published at ALTA 2017. Following that, we extended the model to incorporate sequence information both in sentences and documents, by using a hierarchical and bidirectional long short-term memory (bi-LSTM) architecture. We incorporated a structured loss to encode the task dependencies and observed that the joint model performed better than the separately trained sentence- and document-level models, especially for the sentence-level task under semi-supervised setting where there are more documents with RILE scores than sentence-level labels. The bi-LSTM model performed

better than average word embeddings, which shows that sequential information is useful for this task. Lastly, in order to obtain *left-right* scores which predict expert surveys, we developed an approach using Probabilistic Soft Logic (PSL) [Kimmig et al. (2012); Bach et al. (2017)], which incorporates contextual information, including RILE predictions using our joint-model alongside as predicates. PSL uses first order logic rules as a template language to define a Markov Random Field over random variables with soft-truth values in the interval [0, 1], and inference is framed as a continuous optimization problem. We observed that the output of the PSL model based on contextual dependencies was closer to expert survey scores, than simple RILE predictions. This work was published at NAACL-HLT 2018.

Joint Sentence–Document Model for Manifesto Text Analysis

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

Abstract

Election manifestos document the intentions, motives, and views of political parties. They are often used for analysing party policies and positions on various issues, as well as for quantifying a party's position on the left–right spectrum. In this paper we propose a model for automatically predicting both types of analysis from manifestos, based on a joint sentence–document approach which performs both sentence-level thematic classification and document-level position quantification. Our method handles text in multiple languages, via the use of multilingual vector-space embeddings. We empirically show that the proposed joint model performs better than state-of-art approaches for the document-level task and provides comparable performance for the sentence level task, using manifestos from thirteen countries, written in six different languages.

1 Introduction

Election manifestos are a core artifact in political text analysis. One of the widely used datasets by political scientists is the Comparative Manifesto Project (CMP) dataset, initiated by Volkens et al. (2011), that collects party 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 the sentence-level with one of over fifty fine-grained political themes, divided into 7 coarse-grained topics (see Table 5). These are important because it can be seen as party positions on fine-grained policy themes and also the coded text can be used for various downstream tasks (Lowe et al., 2011). 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, coder preferences towards particular categories might lead to annotation inconsistencies and affect comparability between manifestos annotated by different coders (Mikhaylov et al., 2012). In order to overcome these challenges, fine and coarse-level manifesto sentence classification was addressed using supervised machine learning techniques (Verberne et al., 2014; Zirn et al., 2016). Nonetheless, manually-coded manifestos remain the crucial data source for studies in computational political science (Lowe et al., 2011; Nanni et al., 2016).

Other than the sentence-level labels, the manifesto text also has document-level signals, which quantify its position on the left–right spectrum (Slapin and Proksch, 2008). Though sentence-level classification and document-level quantification tasks are inter-dependent, existing work handles them separately. We instead propose a joint approach to model the two tasks together. Overall, the contributions of this work are as follows:

- we empirically study the utility of multilingual embeddings for cross-lingual manifesto text analysis — at the sentence (for 57-class classification) and document-levels (for RILE score regression)
- we evaluate the effectiveness of modelling the sentence- and document-level tasks together
- we study the value of *country* information used in conjunction with text for the

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin and Julian Brooke. 2017. Joint Sentence-Document Model for Manifesto Text Analysis. In *Proceedings of Australasian Language Technology Association Workshop*, pages 25–33.

document-level regression task.

2 Related Work

The recent adoption of NLP methods has led to significant advances in the field of Computational Social Science (Lazer et al., 2009), including political science (Grimmer and Stewart, 2013). Some popular tasks addressed with political text include: party position analysis (Biessmann, 2016); political leaning categorization (Akoglu, 2014; Zhou et al., 2011); stance classification (Sridhar et al., 2014); identifying keywords, themes & topics (Karan et al., 2016; Ding et al., 2011); emotion analysis (Rheault, 2016); and sentiment analysis (Bakliwal et al., 2013). The source data includes manifestos, political speeches, news articles, floor debates and social media posts.

With the increasing availability of large-scale datasets and computational resources, large-scale comparative political text analysis has gained the attention of political scientists (Lucas et al., 2015). For example, rather than analyzing the political manifestos of a particular party during an election, mining different manifestos across countries over time can provide deeper comparative insights into political change.

Existing classification models, except (Glavaš et al., 2017), utilize discrete representation of text (i.e., bag of words). Also, most of the work analyzes manifesto text at the country level. Recent work has demonstrated the utility of neural embeddings for multi-lingual coarse-level topic classification (7 major categories) over manifesto text (Glavaš et al., 2017). The authors show that multi-lingual embeddings are more effective in the cross-lingual setting, where labeled data is used from multiple languages. In this work, we focus on cross-lingual fine-grained thematic classification (57 categories in total), where we have labeled data for all the languages.

For the document-level quantification task, much work has used label count aggregation of manually-annotated sentences as features (Lowe et al., 2011; Benoit and Däubler, 2014), while other work has used dictionary- based supervised methods, or unsupervised factor analysis based techniques (Hjorth et al., 2015; Bruinsma and Gemenis, 2017). The latter method uses discrete word representations and deals with mono-lingual text only. In Glavas et al. (2017), the authors leverage neural embeddings for cross-lingual EU parliament speech text quantification with two pivot texts for extreme left and right positions. They represent the documents using word embeddings averaged with TF-IDF scores as weights. All these approaches model the sentence and document-level tasks separately.

3 Manifesto Text Analysis

In the CMP, trained annotators manually label manifesto sentences according to the 57 fine-grained political categories (shown in Table 5), which are grouped into seven policy areas: External Relations, Freedom and Democracy, Political System, Economy, Welfare and Quality of Life, Fabric of Society, and Social Groups. Political parties either write their promises as a bulleted list of individual sentences, or structured as paragraphs (an example is given in Figure 4), providing more information on topic coherence. Also the length of documents, measured as the number of sentences, varies greatly between manifestos. The typical length (in sentences) over manifestos (948 in total) from 13 countries — Austria, Australia, Denmark, Finland, France, Germany, Italy, Ireland, New Zealand, South Africa, Switzerland, United Kingdom and United States — is 516.7±667. Variance in the number of sentences across documents in conjunction with class imbalance makes automated thematic classification a challenging task.

While annotating, a sentence is split into multiple segments if it discusses unrelated topics or different aspects of a larger policy, e.g. (as indicated by the different colors, and associated integer labels):

> We need to address our close ties with our neighbours (107) as well as the unique challenges facing small business owners in this time of economic hardship. (402)

Such examples are not common, however.[1] Also the segmentation was shown to be inconsistent and to have no effect on quantifying the proportion of sentences discussing various topics and document-level regression tasks (Däubler et al., 2012). Hence, consistent with previous work

[1] In Däubler et al. (2012), based on a sample of 15 manifestos, the authors noted that around 7.7% of sentences encode multiple topics.

(Biessmann, 2016; Glavaš et al., 2017), we consider the sentence-level classification to be a multi-class single-label problem. We use the segmented text when available (especially for evaluation), and complete sentences otherwise.

A manifesto as a whole can be positioned on the left–right spectrum based on the proportion of topics discussed. We use the RILE score, which is defined as the difference between the count of sentences discussing left- and right-leaning topics (Budge and Laver, 1992):

$$\text{RILE} = \sum_{r \in R} \text{per}_r - \sum_{l \in L} \text{per}_l \quad (1)$$

where R and L denote right and left political themes (see Figure 5), and per_t denotes the share of each topic t as given in Table 5, per document. Note that the RILE score is provided for almost all the manifestos in the CMP dataset, but the sentence-level annotations are provided only for a subset of manifestos. That is, in some cases, the underlying annotations that the RILE score calculation was based on is often not available for a given manifesto.

4 Proposed Approach

We propose a joint sentence–document model to classify manifesto sentences into one out of 57 categories and also quantify the document-level RILE score. The joint formulation is employed not only to capture the task inter-dependencies, but also to use annotations at different levels of granularity (sentence and document) effectively — a RILE score is available for 948 manifestos from 13 countries, whereas sentence-level annotations are available only for 235 manifestos. We use a hierarchical neural network to model the sentence-level classification and document-level regression tasks. The proposed architecture is given in Figure 1. Since the text across countries is multi-lingual in nature, we use multi-lingual embeddings to represent words (e_w) (Ammar et al., 2016). We refer to the total set of manifestos available for training as D, and the subset which is annotated with sentence-level labels as D_s. We denote each manifesto as d, which has l_d sentences $s_1, s_2, ..., s_{l_d}$. We also use i to index documents (i=d) wherever necessary to avoid ambiguity in differentiating from sentence-level variables.

4.1 Sentence-level Model

We represent each sentence using the average embedding of its constituent words, $\boldsymbol{s}_j = \frac{1}{|s_j|} \sum_{w \in s_j} e_w$. The average embedding representation is given as input to a hidden layer with rectified linear activation units (ReLU) to get the hidden representation. Finally, the predictions are obtained using a softmax layer, which takes the hidden representation as input and gives the probability of 57 classes as output, denoted $\hat{\mathbf{y}}_{ij}$. We use the cross-entropy loss function for the sentence-level model. For sentences in D_s, with ground truth labels $\mathbf{y}_{ij}$ (using a one-hot encoding), the loss function is given as follows:

$$\mathcal{L}_S = -\frac{1}{|D_s|} \sum_{i=1}^{|D_s|} \sum_{j=1}^{l_d} \sum_{k=1}^{K} y_{ijk} \log \hat{y}_{ijk} \quad (2)$$

4.2 Joint Sentence–Document Model

Using the hierarchical neural network, we model the sentence-level classification and document-level regression tasks together. In the joint model, we use an unrolled (time-distributed) neural network model for the sentences in a manifesto (d). Here, the model minimizes cross-entropy loss for sentences over each temporal layer ($j = 1 \ldots l_d$). We use average-pooling with the concatenated hidden representations ($\mathbf{h}_{ij}$) and predicted output distributions ($\hat{\mathbf{y}}_{ij}$) of individual sentences, to represent a document,[2] i.e., $\mathbf{r}_d = \frac{1}{|l_d|} \sum_{j \in d} \begin{bmatrix} \hat{\mathbf{y}}_{ij} \\ \mathbf{h}_{ij} \end{bmatrix}$.

The range of RILE is $[-100, 100]$, which we scale to the range $[-1, 1]$. Hence we use a final $\tanh$ layer, with $\hat{z}_i = \mathbf{w}_r^\top \mathbf{h}_d + b$, where $\mathbf{h}_d = \text{ReLU}(W_d^\top \mathbf{r}_d)$. Since it is a regression task, we minimize the mean-squared error loss function between the predicted $\hat{z}_i$ and actual RILE score z_i,

$$\mathcal{L}_D = \frac{1}{|D|} \sum_{i=1}^{|D|} ||\hat{z}_i - z_i||_2^2 \quad (3)$$

Overall, the loss function for the joint model, combining Equations 2 and 3, is:

$$\alpha \mathcal{L}_S + (1 - \alpha) \mathcal{L}_D \quad (4)$$

where $0 \leq \alpha \leq 1$ is a hyperparameter which is tuned on a development set.

We evaluate both cascaded and joint training for this objective function:

[2] We observed that the concatenated representation performed better than using either hidden representation or output distribution.

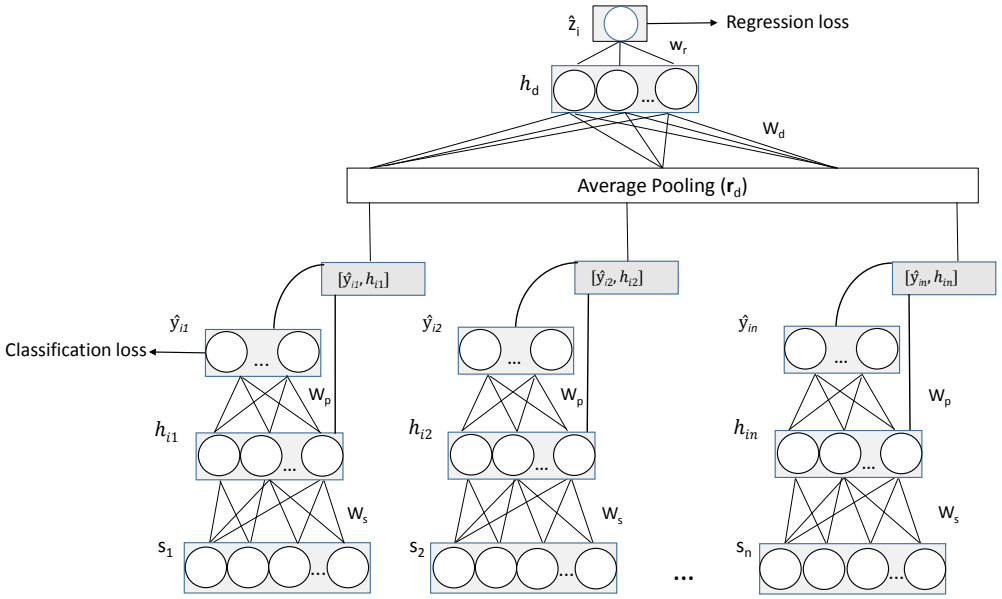


Figure 1: Hierarchical Neural Network for Joint Sentence–Document Analysis. s_1, s_2,...s_n are input sentences (n=l_d), W_s and W_p are shared across unrolled sentences. $\hat{\mathbf{y}}_{ij}$ denotes 57 classes and $\hat{z}_i$ denotes the estimated RILE score

Cascaded Training: The sentence-level model is trained using D_s, to minimize $\mathcal{L}_S$ in Equation 2, and the pre-trained sentence-level model is used to obtain document-level representation $\mathbf{r_d}$ for all the manifestos in the training set D. Then the document-level regression task is trained to minimize $\mathcal{L}_D$ from Equation 3. Here, the sentence-level model parameters are fixed when the document-level regression model is trained using $\mathbf{r_d}$.

Joint Training: The entire network is updated by minimizing the joint loss function from Equation 4. As in cascaded training, the sentence-level model is pre-trained using labeled sentences. Here the sentence-level model uses both labeled and unlabeled data.

We use the Adam optimizer (Kingma and Ba, 2014) for parameter estimation. The proposed architecture evaluates the effectiveness of posing sentence-level topic classification as a precursor to perform document-level RILE prediction, rather than learning a model directly. We also study the effect of the quantity of annotated text at both the sentence- and document-level for the RILE prediction task.

5 Experiments

5.1 Setting

As mentioned earlier, we use manifestos collected and annotated by political scientists as part of CMP. In this work, we used 948 manifestos from 13 countries, which are written in 6 different languages — Danish (Denmark), English (Australia, Ireland, New Zealand, South Africa, United Kingdom, United States), Finnish (Finland), French (France), German (Austria, Germany, Switzerland), and Italian (Italy). Out of the 948 manifestos, 235 are annotated with sentence level labels (from Table 5). We have RILE scores for all the 948 manifestos. Statistics about number of annotated documents and sentences across languages are given in Table 1. Class distribution based on average percentage of sentences coded under each class is given in Figure 2. Top-3 frequent set of classes include 000 (above 8%) , 504 (6-8%) and 305 & 503 (4-6%); and 26 classes occur 0-1%.

We use off-the-shelf pre-trained multi-lingual word embeddings[3] to represent words. We empirically chose embeddings trained using translation invariance approach (Ammar et al., 2016), with size 512 for our work. The neural network model has a single hidden layer for all the sentence and document-level approaches.

[3] http://128.2.220.95/multilingual

Lang.	# Docs (Ann.)	# Sents (Ann.)
Danish	175 (36)	32161 (8762)
English	312 (94)	227769 (73682)
Finnish	97 (16)	18717 (8503)
French	53 (10)	24596 (5559)
German	216 (65)	146605 (79507)
Italian	95 (14)	40010 (4918)
Total	948 (235)	489858 (180931)

Table 1: Statistics of dataset, 'Ann.' refers to annotated at sentence level.

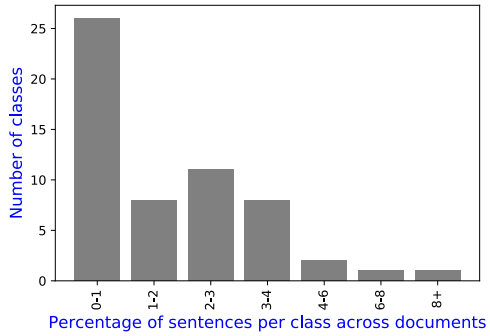


Figure 2: Class distribution based on average percentage of sentences coded under each class

5.2 Sentence-Level Classification

We first compare traditional bag-of-words discrete representation with distributed neural representation for words for fine-grained thematic classification, under mono-lingual training setting (*Mono-lingual).* Hence we compare the following approaches.

Bag-of-words (*BoW-LR,* BoW-NN): We use TF-IDF representation for sentences and build a model for each language separately. We use Logistic Regression classifier (Biessmann, 2016), which is referred as *BoW-LR.* We also use Neural Network classifier, which we refer to as *BoW-NN.*

Language-wise average embedding (*AE-NN$_m$):* We build a neural network classifier per language, with average multi-lingual neural embedding as sentence representation.

Since distributed representation allows to leverage text across languages, we evaluate the following approaches with combined training sentences across languages (*Cross-lingual).*

Convolutional Neural Network (*CNN):* CNN was shown to be effective for cross-lingual manifesto text coarse-level topic classification (Glavaš et al., 2017). So, we evaluate CNN with a similar architecture — single convolution layer (32 filters with window size 3), followed by single max pooling layer and finally a softmax layer. We use multi-lingual neural embeddings to represent words.

Combined average embedding (*AE-NN$_c$):* We build a neural network classifier with training instances combined across languages, with average neural embedding as sentence representation. This is our proposed approach for sentence-level model.

Commonly for all empirical evaluations, we compute micro-averaged performance with 80-20% train-test ratio across 10 runs with random split (at document level), where the 80% split also contains sentence level annotated documents proportionally. Optimal model parameters we found for the proposed model (Figure 1) are $|\mathbf{h}_{ij}| = 300$ (for sentences), $|\mathbf{h}_d| = 10$. We compute F-score[4] to evaluate sentence classification performance. Sentence classification performance is given in Table 2. Under mono-lingual setting (Table 2), using word embeddings did not provide better performance compared to bag-of-words.

Under cross-lingual setting, *AE-NN$_c$* is the sentence-level neural network model. We use *AE-NN$_c$* in the *cascaded training* for obtaining document-level RILE prediction. Note that in *cascaded training*, sentence and document-level models are trained separately in a cascaded fashion. Joint-training results where the sentence model is trained in a semi-supervised way together with document-level regression task is referred to as *JT$_s$*. We set α=0.4 (in equation 4) empirically which gave the best score for both sentence and document-level tasks. We observed a trade-off in performance with different α, with lesser α (0.1), document-level correlation increases (to 0.52) while sentence-level F-score decreases (to 0.33). Higher value of α (0.9) gives performance closer to cascaded training. *JT$_s$* has a comparable performance with *AE-NN$_c$*. The proposed approach (joint-training) does not provide any improvement for the sentence classification task.

[4]Harmonic mean of precision and recall, https://en.wikipedia.org/wiki/F1_score

	Mono-lingual			Cross-lingual		
Lang.	*BoW-LR*	*BoW-NN*	*AE-NN$_m$*	*CNN*	*AE-NN$_c$*	*JT$_s$*
da	0.29	**0.35**	0.24	0.30	0.28	0.30
en	0.36	0.38	**0.42**	0.40	**0.42**	0.41
fi	0.21	0.29	0.26	**0.30**	0.27	0.26
fr	0.28	0.36	0.24	0.36	0.37	**0.38**
de	0.30	0.31	0.31	0.31	0.31	**0.33**
it	0.32	**0.33**	0.25	0.30	0.32	0.26
Avg.	0.32	0.34	0.35	0.34	**0.36**	0.35

Table 2: Micro-Averaged F-measure for sentence classification. Best scores are given in bold.

5.3 Document-Level Regression

For the document-level regression task, the following are baseline approaches. Note that we use $\tanh$ output for all the models, since the range of re-scaled RILE is from -1 to +1.

Bag-of-words (*BoW-NN$_d$):* We use TF-IDF representation for documents and build a neural network model for each language.

Average embedding (*AE-NN$_d$):* We use average embedding of words as document representation to build a neural network model.

Bag-of-Centroids (*BoC):* Here the word embeddings are clustered into K different clusters using K-Means clustering algorithm, and words (1-gram) in each document are assigned to clusters based on its euclidean-distance (dist) to cluster-centroids (C) (Lebret and Collobert, 2014),

$$\text{cluster}(w) = \underset{k}{\text{argmin}}\, \text{dist}(C_k, w).$$

Finally, each document is represented by the distribution of words mapped to different clusters ($1 \times K$ vector). We use a neural network regression model with bag-of-centroids representation. Results with K=1000, which performed best is given in Table 3.

Sentence-level model and RILE formulation (*AE-NN$_c^{rile}$*): Here the predictions of sentence-level model (*AE-NN$_c$*) are used directly with RILE formulation (equation (1)) to derive RILE score for manifestos.

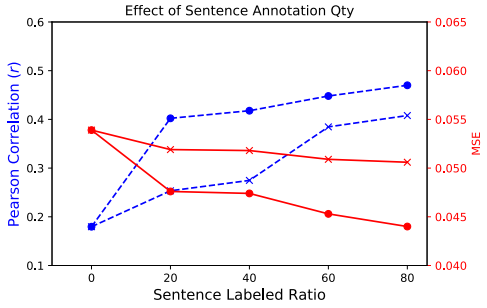
Cross-lingual scaling (*CLS):* This is a recent unsupervised approach for cross-lingual political speech text positioning task (Glavas et al., 2017). Authors use average word-embeddings weighed by TF-IDF score to represent documents.[5] Then a graph is constructed using pair-wise distance of documents. Given two pivots texts for extreme left and right positions [-1, +1], label propagation approach is used to quantify other documents in the graph.

Approach	MSE(↓)	r(↑)
BoW-NN$_d$	0.054	0.23
AE-NN$_d$	0.057	0.14
BoC	0.052	0.33
AE-NN$_c^{rile}$	0.060	0.35
CLS	–	0.24
Cas$_d$	0.050	0.41
JT$_d$	**0.044**	**0.47**

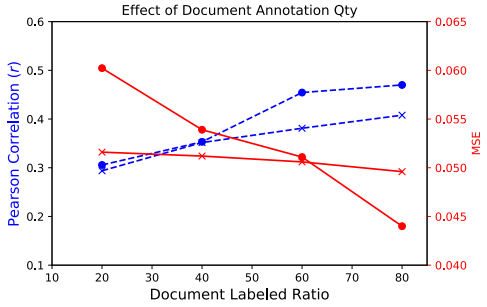
Table 3: RILE score prediction performance. Best scores are given in bold (higher is better for r, and lower is better for MSE).

RILE score regression performance results are given in Table 3. Other than *BoW-NN$_d$* all other approaches are cross-lingual. We evaluate document-level performance using mean-squared-error (MSE) and Pearson correlation (r). Since *CLS* solves it as a classification problem, MSE is not applicable. The proposed approach's performance, using cascaded training is referred to as *Cas$_d$* and jointly trained model is referred to as *JT$_d$*. Overall the jointly trained model performs best for document-level task, with a comparable performance at sentence-level task.

[5] We use this aggregate representation since it was shown to be better than word alignment and scoring approach (Glavas et al., 2017)



(a) Fixing 80% training documents with RILE score, ratio of documents with sentence-level annotations is varied.



(b) Fixing 80% training documents with sentence-level annotations, ratio of documents with RILE score is varied.

Figure 3: Study with Quantity of Annotation. In 3(a) and 3(b) cross(×) denotes Cas_d and circle(○) denotes JT_d

5.4 Quantity of Annotation

We measure the importance of annotated text at sentence and document-level for RILE score regression task. We vary the percentage of labeled data, while keeping the test sample size at 20% as before. In the first setting, we keep the training ratio of documents at 80%, within that 80% we increase the proportion of documents with sentence-level annotations — from 0 (document average embedding setting, *AE-NN$_d$)* to 80%. Results are given in Figure 3a. Similarly, in the other setting, we keep the training set with 80% sentence-level annotated documents (which is ∼20% of the total data), and add documents (with only RILE score), increasing the training set from 20 to 80%. Results of this study are given in Figure 3b. We observed that, jointly-trained model uses sentence-level annotations more effectively than cascaded approach (Figure 3a) — even with less sentence-level annotations. Also, with less document-level signal (up to 40%) for training, both the approaches perform similarly (r). As the training ratio increases, joint-training leverages both sentence and document-level signals effectively.

Approach	MSE	r
stack	0.045 (0.001 ↓)	0.49 (0.02 ↑)
non-linear stack	0.048 (0.004 ↓)	0.48 (0.01 ↑)

Table 4: RILE score prediction performance with *country* information. Difference compared to JT_d is given within paranthesis. ↑ – improvement, ↓ – decrease in performance

5.5 Use of Country Information

Since the definition of left–right varies between countries, we study the influence of *country* information in the proposed model with *joint-training*. We use two ways to incorporate country information (Hoang et al., 2016): (a) *stack* — one-hot encoding (13 countries, 1×13 vector) of each manifesto's *country* is concatenated with hidden representation of the document ($\mathbf{r}_d$ in Figure 1) (b) *non-linear stack* — one-hot-encoded country vector is passed through a hidden layer with $\tanh$ non-linear activation and concatenated with $\mathbf{r}_d$. With both the models we observed mild improvement in correlation (given in Table 4).

6 Conclusion and Future Work

In this work we evaluated the utility of a joint sentence–document model for sentence-level thematic classification and document-level RILE score regression tasks. Our observations are as follows: (a) joint model performs better than state-of-art approaches for document-level regression task (b) joint-training leverages sentence-level annotations more effectively than cascaded approach for RILE score regression task, with no gains for sentence classification task. There are many extensions possible to the current work. First is to handle class imbalance in the dataset with a cost-sensitive objective function. Secondly, CNN gave a comparable performance with Neural Network, which motivates the need to evaluate an end-end sequential architecture to obtain sentence and document embeddings. Off-the-shelf embeddings leads to out-of-vocabulary scenarios. It could be beneficial to adapt word-embeddings with manifesto corpus. Finally, background information such as country can be leveraged more effectively.

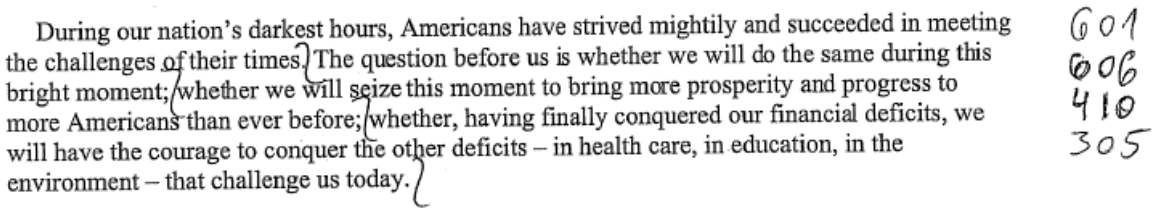


Figure 4: Manifesto snippet for Democratic Party of USA, 2000 — $\int$ denotes sentence segment. See Table 5 for code description.

CMP Coding Scheme	
• Domain 1. External Relations	411 Technology and Infrastructure: Positive
101 Foreign Special Relationships: Positive	412 Controlled Economy
102 Foreign Special Relationships: Negative	413 Nationalisation
103 Anti-Imperialism	414 Economic Orthodoxy
104 Military: Positive	415 Marxist Analysis
105 Military: Negative	416 Anti-Growth Economy: Positive
106 Peace	
107 Internationalism: Positive	• Domain 5: Welfare and Quality of Life
108 European Community/Union: Positive	501 Environmental Protection
109 Internationalism: Negative	502 Culture: Positive
110 European Community/Union: Negative	503 Equality: Positive
	504 Welfare State Expansion
• Domain 2: Freedom and Democracy	505 Welfare State Limitation
201 Freedom and Human Rights	506 Education Expansion
202 Democracy	507 Education Limitation
203 Constitutionalism: Positive	
204 Constitutionalism: Negative	• Domain 6: Fabric of Society
	601 National Way of Life: Positive
• Domain 3: Political System	602 National Way of Life: Negative
301 Decentralisation	603 Traditional Morality: Positive
302 Centralisation	604 Traditional Morality: Negative
303 Governmental and Administrative Efficiency	605 Law and Order: Positive
304 Political Corruption	606 Civic Mindedness: Positive
305 Political Authority	607 Multiculturalism: Positive
	608 Multiculturalism: Negative
• Domain 4: Economy	
401 Free Market Economy	• Domain 7: Social Groups
402 Incentives: Positive	701 Labour Groups: Positive
403 Market Regulation	702 Labour Groups: Negative
404 Economic Planning	703 Agriculture and Farmers: Positive
405 Corporatism/Mixed Economy	704 Middle Class and Professional Groups
406 Protectionism: Positive	705 Underprivileged Minority Groups
407 Protectionism: Negative	706 Non-economic Demographic Groups
408 Economic Goals	
409 Keynesian Demand Management	
410 Economic Growth: Positive	000 No meaningful category applies

Table 5: Comparative Manifesto Project — 57 policy themes. *Left* topics are given in *red* and *right* topics are given in *blue* and the *rest* are considered neutral

References

Leman Akoglu. 2014. Quantifying political polarity based on bipartite opinion networks. In *AAAI*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925* .

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. In *ACL*.

Kenneth Benoit and Thomas Däubler. 2014. Putting text in context: How to estimate better left-right positions by scaling party manifesto data using item response theory. In *Mapping Policy Preferences from Texts Conference*.

Felix Biessmann. 2016. Automating political bias prediction. In *arXiv:1608.02195*.

B. Bruinsma and K. Gemenis. 2017. Validating Wordscores. *ArXiv e-prints* .

Ian Budge and Michael Laver. 1992. *Party policy and government coalitions*. St. Martin's Press New York.

Thomas Däubler, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2012. Natural sentences as valid units for coded political texts. *British Journal of Political Science* .

Zhuoye Ding, Qi Zhang, and Xuanjing Huang. 2011. Keyphrase extraction from online news using binary integer programming. In *IJCNLP*.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of topics in political texts. In *ACL WS*.

Goran Glavas, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In *EACL*.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* .

Frederik Hjorth, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. 2015. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics* .

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Incorporating side information into recurrent neural network language models. In *NAACL-HLT*.

Mladen Karan, Jan Šnajder, Daniela Širinic, and Goran Glavaš. 2016. Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *LaTeCH*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323(5915):721.

Rémi Lebret and Ronan Collobert. 2014. N-gram-based low-dimensional representation for document classification. *arXiv preprint arXiv:1412.6277* .

Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling policy preferences from coded political texts. In *Legislative studies quarterly*. Wiley Online Library.

Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. *Political Analysis* 23(2):254–277.

Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. In *Political Analysis*.

Federico Nanni, Cäcilia Zirn, Goran Glavaš, Jason Eichorst, and Simone Paolo Ponzetto. 2016. Topfish: topic-based analysis of political position in us electoral campaigns. In *PolText*.

Ludovic Rheault. 2016. Expressions of anxiety in political texts. In *NLP+ CSS 2016*. page 92.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. Wiley Online Library, volume 52, pages 705–722.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *ACL*.

Suzan Verberne, Eva D'hondt, Antal van den Bosch, and Maarten Marx. 2014. Automatic thematic classification of election manifestos. In *Information Processing & Management*.

Andrea Volkens, Onawa Lacewell, and Sven Regel Henrike Schultzeand Annika Werner. Pola Lehmann. 2011. The manifesto data collection. manifesto project (mrg/cmp/marpor). In *Wissenschaftszentrum Berlin fur Sozialforschung (WZB)*.

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. 2011. Classifying the political leaning of news articles and users from user votes. In *ICWSM*.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos. In *PolText*.

Hierarchical Structured Model for Fine-to-coarse Manifesto Text Analysis

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

Election manifestos document the intentions, motives, and views of political parties. They are often used for analysing a party's fine-grained position on a particular issue, as well as for coarse-grained positioning of a party on the left–right spectrum. In this paper we propose a two-stage model for automatically performing both levels of analysis over manifestos. In the first step we employ a hierarchical multi-task structured deep model to predict fine- and coarse-grained positions, and in the second step we perform post-hoc calibration of coarse-grained positions using probabilistic soft logic. We empirically show that the proposed model outperforms state-of-art approaches at both granularities using manifestos from twelve countries, written in ten different languages.

1 Introduction

The adoption of NLP methods has led to significant advances in the field of computational social science (Lazer et al., 2009), including political science (Grimmer and Stewart, 2013). Among a myriad of data sources, election manifestos are a core artifact in political analysis. One of the most widely used datasets by political scientists is the Comparative Manifesto Project (CMP) dataset (Volkens et al., 2017), which contains manifestos in various languages, covering over 1000 parties across 50 countries, from elections dating back to 1945.

In CMP, a subset of the manifestos has been manually annotated at the sentence-level with one of 57 political themes, divided into 7 major categories.[1] Such categories capture party positions (FAVORABLE, UNFAVORABLE or NEITHER) on fine-grained policy themes, and are also useful for downstream tasks including calculating manifesto-level (policy-based) left–right position scores (Budge et al., 2001; Lowe et al., 2011; Däubler and Benoit, 2017). An example sentence from the Green Party of England and Wales 2015 election manifesto where they take an UNFAVORABLE position on MILITARY is:

> *We would: Ensure that ... less is spent on military research.*

Elsewhere, they take a FAVORABLE position on WELFARE STATE:

> *Double Child Benefit.*

Such manual annotations are labor-intensive and prone to annotation inconsistencies (Mikhaylov et al., 2012). In order to overcome these challenges, supervised sentence classification approaches have been proposed (Verberne et al., 2014; Subramanian et al., 2017).

Other than the sentence-level labels, the manifesto text also has a document-level score that quantifies its position on the left–right spectrum. Different approaches have been proposed to derive this score, based on alternate definitions of "left–right" (Slapin and Proksch, 2008; Benoit and Laver, 2007; Lo et al., 2013; Däubler and Benoit, 2017). Among these, 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 RIGHT and LEFT positions on (pre-determined) policy themes across sentences in a manifesto (Volkens et al., 2013); for instance, UNFAVORABLE position on MILITARY is categorized as LEFT. RILE is popular in CMP in particular, as mapping individual sentences to LEFT/RIGHT/NEUTRAL categories has

[1] https://manifesto-project.wzb.eu/coding_schemes/mp_v5

Proceedings of NAACL-HLT 2018, pages 1964–1974
New Orleans, Louisiana, June 1 - 6, 2018.

been shown to be less sensitive to systematic errors than other sentence-level class sets (Klingemann et al., 2006; Volkens et al., 2013).

Finally, expert survey scores are gaining popularity as a means of capturing manifesto-level political positions, and are considered to be context- and time-specific, unlike RILE (Volkens et al., 2013; Däubler and Benoit, 2017). We use the Chapel Hill Expert Survey (CHES) (Bakker et al., 2015), which comprises aggregated expert surveys on the ideological position of various political parties. Although CHES is more subjective than RILE, the CHES scores are considered to be the gold-standard in the political science domain.

In this work, we address both fine- and coarse-grained multilingual manifesto text policy position analysis, through joint modeling of sentence-level classification and document-level positioning (or ranking) tasks. We employ a two-level structured model, in which the first level captures the structure within a manifesto, and the second level captures context and temporal dependencies across manifestos. Our contributions are as follows:

- we employ a hierarchical sequential deep model that encodes the structure in manifesto text for the sentence classification task;
- we capture the dependency between the sentence- and document-level tasks, and also utilize additional label structure (categorization into LEFT/RIGHT/NEUTRAL: Volkens et al. (2013)) using a joint-structured model;
- we incorporate contextual information (such as political coalitions) and encode temporal dependencies to calibrate the coarse-level manifesto position using probabilistic soft logic (Bach et al., 2015), which we evaluate on the prediction of the RILE index or expert survey party position score.

2 Related Work

Analysing manifesto text is a relatively new application at the intersection of political science and NLP. One line of work in this space has been on sentence-level classification, including classifying each sentence according to its major political theme (1-of-7 categories) (Zirn et al., 2016; Glavaš et al., 2017a), its position on various policy themes (Verberne et al., 2014; Biessmann, 2016; Subramanian et al., 2017), or its relative disagreement with other parties (Menini et al., 2017). Recent approaches (Glavaš et al., 2017a; Subramanian et al., 2017) have also handled multilingual manifesto text (given that manifestos span multiple countries and languages; see Section 5.1) using multilingual word embeddings.

At the document level, there has been work on using label count aggregation of (manually-annotated) fine-grained policy positions, as features for inductive analysis (Lowe et al., 2011; Däubler and Benoit, 2017). Text-based approaches has used dictionary-based supervised methods, unsupervised factor analysis based techniques and graph propagation based approaches (Hjorth et al., 2015; Bruinsma and Gemenis, 2017; Glavaš et al., 2017b). A recent paper closely aligned with our work is Subramanian et al. (2017), who address both sentence- and document-level tasks jointly in a multilingual setting, showing that a joint approach outperforms previous approaches. But they do not exploit the structure of the text and use a much simpler model architecture: averages of word embeddings, versus our bi-LSTM encodings; and they do not leverage domain information and temporal regularities that can influence policy positions (Greene, 2016). This work will act as a baseline in our experiments in Section 5.

Policy-specific position classification can be seen as related to target-specific stance classification (Mohammad et al., 2017), except that the target is not explicitly mentioned in most cases. Secondly, manifestos have both fine- and coarse-grained positions, similar to sentiment analysis (McDonald et al., 2007). Finally, manifesto text is well structured within and across documents (based on coalition), has temporal dependencies, and is multilingual in nature.

3 Proposed Approach

In this section, we detail the first step of our two-stage approach. We use a hierarchical bidirectional long short-term memory ("bi-LSTM") model (Hochreiter and Schmidhuber, 1997; Graves et al., 2013; Li et al., 2015) with a multi-task objective for the sentence classification and document-level regression tasks. A post-hoc calibration of coarse-grained manifesto position is given in Section 4.

Let D be the set of manifestos, where a manifesto $d \in D$ is made up of L sentences, and a sentence s_i has T words: $w_{i1}, w_{i2}, ...w_{iT}$. The set $D_s \subset D$ is annotated at the sentence-level

with positions on fine-grained policy issues (57 classes). The task here is to learn a model that can: (a) classify sentences according to policy issue classes; and (b) score the overall document on the policy-based left–right spectrum (RILE), in an inter-dependent fashion.

Word encoder: We initialize word vector representations using a multilingual word embedding matrix, W_e. We construct W_e by aligning the embedding matrices of all the languages to English, in a pair-wise fashion. Bilingual projection matrices are built using pre-trained FastText monolingual embeddings (Bojanowski et al., 2017) and a dictionary D constructed by translating 5000 frequent English words using Google Translate. Given a pair of embedding matrices E (English) and O (Other), we use singular value decomposition of $O^T DE$ (which is $U\Sigma V^T$) to get the projection matrix (W^*=UV^T), since it also enforces monolingual invariance (Artetxe et al., 2016; Smith et al., 2017). Finally, we obtain the aligned embedding matrix, W_e, as OW^*.

We use a bi-LSTM to derive a vector representation of each word in context. The bi-LSTM traverses the sentence s_i in both the forward and backward directions, and the encoded representation for a given word $\mathbf{w}_{it} \in s_i$, is defined by concatenating its forward ($\overrightarrow{\mathbf{h}}_{it}$) and backward hidden states ($\overleftarrow{\mathbf{h}}_{it}$), $t \in [1, T]$.

Sentence model: Similarly, we use a bi-LSTM to generate a sentence embedding from the word-level bi-LSTM, where each input sentence s_i is represented using the last hidden state of both the forward and backward LSTMs. The sentence embedding is obtained by concatenating the hidden representations of the sentence-level bi-LSTM, in both the directions, $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{h}}_i]$, i $\in [1, L]$. With this representation, we perform fine-grained classification (to one-of-57 classes), using a softmax output layer for each sentence. We minimize the cross-entropy loss for this task, over the sentence-level labeled set $D_s \subset D$. This loss is denoted $\mathcal{L}_S$.

Document model: To represent a document d we use average-pooling over the sentence representations $\mathbf{h}_i$ and predicted output distributions ($\mathbf{y}_i$) of individual sentences,[2] i.e., $\mathbf{V}_d = \frac{1}{L}\sum_{i \in d} \begin{bmatrix} \mathbf{y}_i \\ \mathbf{h}_i \end{bmatrix}$. The range of RILE is $[-100, 100]$, which we scale to the range $[-1, 1]$, and model using a final $\tanh$ layer. We minimize the mean-squared error loss function between the predicted $\hat{r}_d$ and actual RILE score r_d, which is denoted as $\mathcal{L}_D$:

$$\mathcal{L}_D = \frac{1}{|D|}\sum_{d=1}^{|D|} \|\hat{r}_d - r_d\|_2^2 \quad (1)$$

Overall, the loss function for the joint model (Figure 1), combining $\mathcal{L}_S$ and $\mathcal{L}_D$, is:

$$\mathcal{L}_J = \alpha\mathcal{L}_S + (1-\alpha)\mathcal{L}_D \quad (2)$$

where $0 \leq \alpha \leq 1$ is a hyper-parameter which is tuned on a development set.

3.1 Joint-Structured Model

The RILE score is calculated directly from the sentence labels, based on mapping each label according to its positioning on policy themes, as LEFT, RIGHT and NEUTRAL (Volkens et al., 2013). Specifically, 13 out of 57 classes are categorized as LEFT, another 13 as RIGHT, and the rest as NEUTRAL. We employ an explicit structured loss which minimizes the deviation between sentence-level LEFT/RIGHT/NEUTRAL polarity predictions $\mathbf{p}$ and the document-level RILE score. The motivation to do this is two-fold: (a) enabling interaction between the sentence- and document-level tasks with homogeneous target space (polarity and RILE); and (b) since we have more documents with just RILE and no sentence-level labels,[3] augmenting an explicit semi-supervised learning objective could propagate down the RILE label to generate sentence labels that concord with the document score.

For the sentence-level polarity prediction (shown in Figure 1), we use cross-entropy loss over the sentence-level labeled set $D_s \subset D$, which is denoted as $\mathcal{L}_{S_P}$. The explicit structured sentence-document loss is given as:

$$\mathcal{L}_{struc} = \frac{1}{|D|}\sum_{d=1}^{|D|}\left(\frac{1}{L_d}\sum_{i \in d}(p_{i_{right}} - p_{i_{left}}) - r_d\right)^2 \quad (3)$$

[2] Preliminary experiments suggested that this representation performs better than using either hidden representations or just the output distribution.

[3] Strictly speaking, for these documents even, sentence annotation was used to derive the RILE score, but the sentence-level labels were never made available.

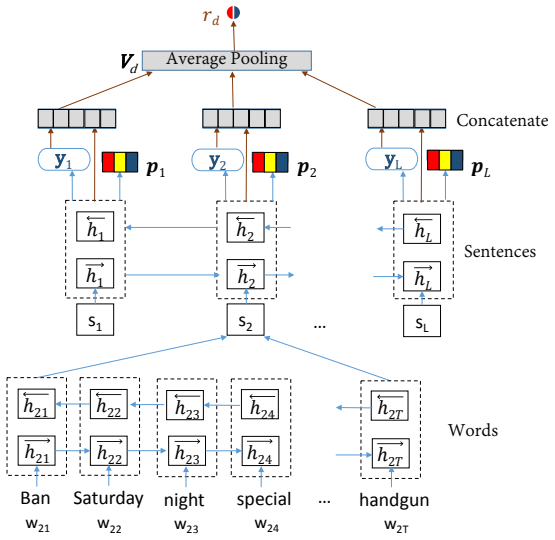


Figure 1: Hierarchical bi-LSTM for joint sentence–document analysis ($\mathbf{y}_i$ denotes the predicted 57-class distribution of sentence s_i; $\mathbf{p}_i$ denotes the distribution over LEFT (in red), RIGHT (in blue) and NEUTRAL (in yellow); r_d denotes the RILE score of d).

where $p_{i_{right}}$ and $p_{i_{left}}$ are the predicted RIGHT and LEFT class probabilities for a sentence s_i ($\in d$), r_d is the actual RILE score for the document d, and L_d is the length of each document, d $\in$ D. We augment the joint model's loss function (Equation (2)) with $\mathcal{L}_{S_P}$ and $\mathcal{L}_{struc}$ to generate a regularized multi-task loss:

$$\mathcal{L}_T = \mathcal{L}_J + \beta\mathcal{L}_{S_P} + \gamma\mathcal{L}_{struc} \quad (4)$$

where $\beta, \gamma \geq 0$ are hyper-parameters which are, once again, tuned on the development set. We refer to the model trained with Equation (2) as "*Joint*", and that trained with Equation (4) as "*Joint*$_{struc}$".

4 Manifesto Position Re-ranking

We leverage party-level information to enforce smoothness and regularity in manifesto positioning on the left–right spectrum (Greene, 2016). For example, manifestos released by parties in a coalition are more likely to be closer in RILE score, and a party's position in an election is often a relative shift from its position in earlier election, so temporal information can provide smoother estimations.

4.1 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). HL-MRFs have been used for many tasks including political framing analysis on Twitter (Johnson et al., 2017) and user stance classification on socio-political issues (Sridhar et al., 2014). These models can be specified using Probabilistic Soft Logic ("PSL") (Bach et al., 2015), a weighted first order logical template language. An example of a PSL rule is

$$\lambda : \texttt{P(a)} \wedge \texttt{Q(a,b)} \rightarrow \texttt{R(b)}$$

where P, Q, and R are predicates, a and 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:

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\left(-\sum_{r=1}^{M} \lambda_r \phi_r(\mathbf{Y}, \mathbf{X})\right), \quad (5)$$
$$\phi_r(\mathbf{Y}, \mathbf{X}) = \max\{l_r(\mathbf{Y}, \mathbf{X}), 0\}^{\rho_r},$$

where $\phi_r(\mathbf{Y}, \mathbf{X})$ is a hinge-loss potential corresponding to an instantiation of a rule, and is specified by a linear function l_r and optional exponent $\rho_r \in \{1, 2\}$. Note that the hinge-loss potential captures the distance to satisfaction.[4]

4.2 PSL Model

Here we elaborate our PSL model (given in Table 1) based on coalition information, manifesto content-based features (manifesto similarity and right–left ratio), and temporal dependency. Our target pos (calibrated RILE) is a continuous variable $[0, 1]$, where 1 indicates that a manifesto occupies an extreme right position, 0 denotes an extreme left position, and 0.5 indicates center. Each instance of a manifesto and its party affiliation are denoted by the predicates Manifesto and Party.

Coalition: We model multi-relational networks based on regional coalitions within a given country (RegCoalition),[5] and also cross-country coalitions in the European parliament

[4] Degree of satisfaction for the example PSL rule r, $\neg$P $\vee$ $\neg$Q $\vee$ R, using the Lukasiewicz co-norm is given as $\min\{2 - \texttt{P} - \texttt{Q} + \texttt{R}, 1\}$. From this, the distance to satisfaction is given as $\max\{\texttt{P} + \texttt{Q} - \texttt{R} - 1, 0\}$, where $\texttt{P} + \texttt{Q} - \texttt{R} - 1$ indicates the linear function l_r.

[5] http://www.parlgov.org/

PSL$_{coal}$ — Coalition features
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{SameElec(x,y)} \wedge \texttt{RegCoalition(a,b)} \wedge \texttt{pos(x)} \rightarrow \texttt{pos(y)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{SameElec(x,y)} \wedge \texttt{RegCoalition(a,b)} \wedge \neg\texttt{pos(x)} \rightarrow \neg\texttt{pos(y)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{Recent(x,y)} \wedge \texttt{EUCoalition(a,b)} \wedge \texttt{pos(x)} \rightarrow \texttt{pos(y)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{Recent(x,y)} \wedge \texttt{EUCoalition(a,b)} \wedge \neg\texttt{pos(x)} \rightarrow \neg\texttt{pos(y)}$
Transitivity
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{Manifesto(z)} \wedge \texttt{Party(z,c)} \wedge \texttt{SameElec(x,y)} \wedge \texttt{SameElec(y,z)} \wedge \texttt{RegCoalition(a,b)} \wedge \texttt{RegCoalition(b,c)} \wedge \texttt{pos(x)} \rightarrow \texttt{pos(z)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{Manifesto(z)} \wedge \texttt{Party(z,c)} \wedge \texttt{SameElec(x,y)} \wedge \texttt{SameElec(y,z)} \wedge \texttt{RegCoalition(a,b)} \wedge \texttt{RegCoalition(b,c)} \wedge \neg\texttt{pos(x)} \rightarrow \neg\texttt{pos(z)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{Manifesto(z)} \wedge \texttt{Party(z,c)} \wedge \texttt{Recent(x,y)} \wedge \texttt{Recent(y,z)} \wedge \texttt{EUCoalition(a,b)} \wedge \texttt{EUCoalition(b,c)} \wedge \texttt{pos(x)} \rightarrow \texttt{pos(z)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y,b)} \wedge \texttt{Manifesto(z)} \wedge \texttt{Party(z,c)} \wedge \texttt{Recent(x,y)} \wedge \texttt{Recent(y,z)} \wedge \texttt{EUCoalition(a,b)} \wedge \texttt{EUCoalition(b,c)} \wedge \neg\texttt{pos(x)} \rightarrow \neg\texttt{pos(z)}$
PSL$_{esim}$ — Similarity-based relational feature
$\texttt{Manifesto(x)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Similarity(x,y)} \wedge \texttt{Recent(x,y)} \wedge \texttt{pos(x)} \rightarrow \texttt{pos(y)}$
$\texttt{Manifesto(x)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Similarity(x,y)} \wedge \texttt{Recent(x,y)} \wedge \neg\texttt{pos(x)} \rightarrow \neg\texttt{pos(y)}$
PSL$_{ploc}$ — Right–left ratio
$\texttt{Manifesto(x)} \wedge \texttt{LwRightLeftRatio(x)} \rightarrow \texttt{pos(x)}$
$\texttt{Manifesto(x)} \wedge \neg\texttt{LwRightLeftRatio(x)} \rightarrow \neg\texttt{pos(x)}$
PSL$_{temp}$— Temporal Dependency
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{PreviousManifesto(x,a,t)} \wedge \texttt{pos(t)} \rightarrow \texttt{pos(x)}$
$\texttt{Manifesto(x)} \wedge \texttt{Party(x,a)} \wedge \texttt{PreviousManifesto(x,a,t)} \wedge \neg\texttt{pos(t)} \rightarrow \neg\texttt{pos(x)}$

Table 1: PSL Model: Values for `Similarity`, `LwRightLeftRatio` and `pos` are obtained from the joint-structured model (Figure 1). Except for `pos`, other values are fixed in the network. Domain (`y`) for `SameElec(x, y)` is within the country, and for `Recent(x, y)` covers all the countries. $\neg$ denotes negation. Distance to satisfaction for each ground rule is obtained using a hinge-loss potential, which is then used inside the HL-MRF model (Equation (5)), where `pos` is $\mathbf{Y}$.

(`EUCoalition`).[6] We set the scope of interaction between manifestos (`x` and `y`) from a country to the same election (`SameElec`). For manifestos across countries, we consider only the most recent manifesto (`Recent`) from each party (`y`), released within 4 years relative to `x`. We use a logistic transformation of the number of times two parties have been in a coalition in the past (to get a value between 0 and 1), for both `RegCoalition` and `EUCoalition`. We also construct rules based on transitivity for both the relational features, i.e., parties which have had common coalition partners, even if they were not allies themselves, are likely to have similar policy positions.

Manifesto similarity: Manifestos that are similar in content are expected to have similar RILE scores (and associated sentence-level label distributions), similar to the modeling intuition captured by Burford et al. (2015) in the context of congressional debate vote prediction. For a pair of recent manifestos (`Recent`) we use the cosine similarity (`Similarity`) between their respective document vectors $\mathbf{V_d}$ (Figure 1).

Right–left ratio: For a given manifesto, we compute the ratio of sentences categorized under RIGHT to OTHERS ($\frac{\#\text{ RIGHT}}{\#\text{ RIGHT}+\#\text{ LEFT}+\#\text{ NEUTRAL}}$), where the categorization for sentences is obtained using the joint-structured model (Equation (4)). We also encode the location of sentence l_s in a document, by weighing the count of sentences for each class C by its location value $\sum_{s \in C} \log(l_s + 1)$ (referred to as *loc_lr*). The intuition here is that the beginning parts of a manifesto tends to contain generic information such as preamble, compared to

[6] http://www.europarl.europa.eu

later parts which are more policy-dense. We perform a logistic transformation of *loc_lr* to derive the `LwRightLeftRatio`.

Temporal dependency: We capture the temporal dependency between a party's current manifesto position and its previous manifesto position (`PreviousManifesto`).

Other than for the look-up based random variables, the network is instantiated with predictions (for `Similarity`, `LwRightLeftRatio` and `pos`) from the joint-structured model (Figure 1). All the random variables, except `pos` (which is the target variable), are fixed in the network. These values are then used inside a PSL model for collective probabilistic reasoning, where the first-order logic given in Table 1 is used to define the graphical model (HL-MRF) over the random variables detailed above. Inference on the HL-MRF is used to obtain the most probable interpretation such that it satisfies most ground rule instances, i.e., considering the relational and temporal dependencies.

5 Evaluation

5.1 Experimental Setup

As our dataset, we use manifestos from CMP for European countries only, as in Section 5.5 we will validate the manifesto's overall position on the left-right spectrum, using the Chapel Hill Expert Survey (CHES), which is only available for European countries (Bakker et al., 2015). In this, we sample 1004 manifestos from 12 European countries, written in 10 different languages — Danish (Denmark), Dutch (Netherlands), English (Ireland, United Kingdom), Finnish (Finland), French (France), German (Austria, Germany), Italian (Italy), Portuguese (Portugal), Spanish (Spain), and Swedish (Sweden). Out of the 1004 manifestos, 272 are annotated with both sentence-level labels and RILE scores, and the remainder only have RILE scores (see Table 2 for further statistics).

There are (less) scenarios where a natural sentence is segmented into sub-sentences and annotated with different classes (Däubler et al., 2012). Hence we use NLTK sentence tokenizer followed by heuristics from Däubler et al. (2012) to obtain sub-sentences. Consistent with previous work (Subramanian et al., 2017), we present results with manually segmented and annotated test documents.

Lang.	# Docs (Anntd.)	# Sents (Anntd.)
Danish	175 (36)	29694 (8762)
Dutch	107 (48)	132524 (70559)
English	117 (27)	86603 (34512)
Finnish	97 (16)	17979 (8503)
French	53 (10)	22747 (5559)
German	117 (46)	111376 (73652)
Italian	98 (15)	41455 (5154)
Portuguese	60 (9)	40922 (11077)
Spanish	85 (50)	145355 (93964)
Swedish	95 (15)	19551 (7938)
Total	1004 (272)	648206 (319680)

Table 2: Statistics of dataset ("Anntd." refers to the number of documents with sentence annotations in the second column, and the number of sentences with annotations in the third column).

5.2 Baseline Approaches

Sentence-level baseline approaches include:

- ***BoW-NN***: TF-IDF-weighted unigram bag-of-words representation of sentences (Biessmann, 2016), and monolingual training using a multi-layer perceptron ("MLP") model.
- ***BoT-NN***: Similar to above, but trigram bag-of-words.
- ***AE-NN***: MLP model with average multilingual word embeddings as the sentence representation (Subramanian et al., 2017).
- ***CNN***: Convolutional neural network ("CNN": Glavaš et al. (2017a)) with multilingual word embeddings.
- ***Bi-LSTM***: Simple bi-LSTM over multilingual word embeddings, last hidden units are concatenated to form the sentence representation, and fed directly into a softmax sentence-level layer. We evaluate two scenarios: (1) with a trainable embedding matrix W_e (***Bi-LSTM(+up)***); and (2) without a trainable W_e.

Document-level baseline approaches include:

- ***BoC***: Bag-of-centroids (BoC) document representation based on clustering the word embeddings (Lebret and Collobert, 2014), fed into a neural network regression model.
- ***HCNN***: Hierarchical CNN, where we encode both the sentence and document using stacked CNN layers.

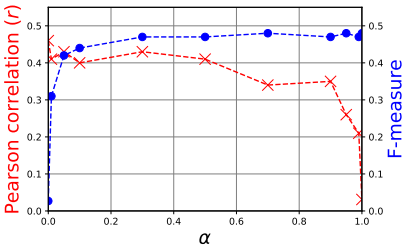
- ***HNN***: State-of-the-art hierarchical neural network model of Subramanian et al. (2017), based on average embedding representations for sentences and the document.

We present results evaluated under two different settings: (a) 80–20% random split averaged across 10 runs to validate the hierarchical model (Section 5.3 and Section 5.4); and (b) temporal setting, where train- and test-set are split chronologically, to validate both the hierarchical deep model and the PSL approach especially, since we encode temporal dependencies (Section 5.5).

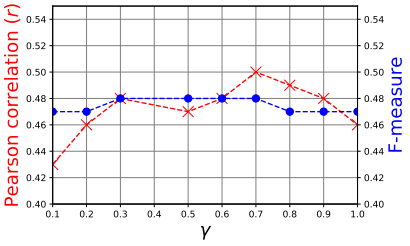
5.3 Hierarchical Sentence- and Document-level Model

We present sentence-level results with a 80–20% random split in Table 3, stratified by country, averaged across 10 runs. For *Bi-LSTM*, we found the setting with a trainable embedding matrix (*Bi-LSTM(+up)*) to perform better than the non-trainable case (*Bi-LSTM*). Hence we use a similar setting for *Joint* and $Joint_{struc}$. We show the effect of α (from Equation (2)) in Figure 2a, based on which we set $\alpha = 0.3$ hereafter. With the chosen model, we study the effect of the structured loss (Equation (4)), by varying γ with fixed $\beta = 0.1$, as shown in Figure 2b. We observe that $\gamma = 0.7$ gives the best performance, and varying β with γ at 0.7 does not result in any further improvement (see Figure 2c). Sentence-level results measured using F-measure, for baseline approaches and the proposed models selected from Figure 2a (*Joint*), Figures 2b and 2c ($Joint_{struc}$) are given in Table 3. We also evaluate the special case of $\alpha = 1$, in the form of sentence-only model $Joint_{sent}$. For the document-level task, results for overall manifesto positioning measured using Pearson's correlation (r) and Spearman's rank correlation (ρ) are given in Table 4. We also evaluate the hierarchical bi-LSTM model with document-level objective only, $Joint_{doc}$.

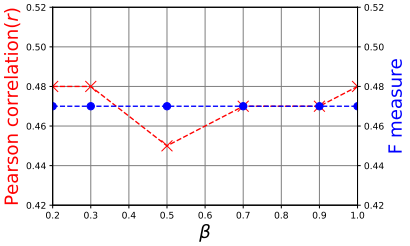
We observe that hierarchical modeling ($Joint_{sent}$, *Joint* and $Joint_{struc}$) gives the best performance for sentence-level classification for all the languages except Portuguese, on which it performs slightly worse than *Bi-LSTM(+up)*. Also, $Joint_{struc}$, does not improve over $Joint_{sent}$. We perform further analysis to see the effect of joint-structured model on the sentence-level task under sparsely-labeled conditions in Section 5.4. On the other hand, for the document-level task,



(a) Effect of α in equation 2.



(b) Effect of γ in equation 4.



(c) Effect of β in equation 4.

Figure 2: Effect of hyper-parameters on sentence- and document-level performance. • denotes F-measure (right axis) and × denotes Pearson correlation (left axis).

the joint model (*Joint*) performs better than $Joint_{doc}$ and all the baseline approaches. Lastly, the joint-structured model ($Joint_{struc}$) provides further improvement over *Joint*.

5.4 Analysis of Joint-Structured Model for Sentence-level task

To understand the utility of joint modeling, especially given that there are more manifestos with document-level labels only than both sentence- and document-level labels, we compare the following two settings: (1) $Joint_{struc}$, which uses additional manifestos with document-level supervision (RILE); and (2) $Joint_{sent}$, which uses manifestos with sentence-level supervision only. We vary the proportion of labeled documents at the sentence-level, from 10% to 80%, to study the effect under sparsely-labeled conditions. Note that 80% is the maximum labeled training data under the cross-validation setting. In other cases, a subset (say 10%) is randomly sampled for train-

Lang.	*BoW-NN*	*BoT-NN*	*AE-NN*	*CNN*	*Bi-LSTM*	*Bi-LSTM(+up)*	*Joint$_{sent}$*	*Joint*	*Joint$_{struc}$*
Danish	0.35	0.33	0.35	0.31	0.38	0.38	**0.44**	0.40	0.43
Dutch	0.41	0.41	0.40	0.34	0.39	0.43	**0.52**	0.50	0.50
English	0.39	0.43	0.43	0.40	0.45	0.47	0.49	**0.50**	0.49
Finnish	0.30	0.34	0.33	0.30	0.38	0.39	**0.44**	0.41	0.42
French	0.36	0.37	0.36	0.37	0.42	0.44	0.48	**0.49**	0.48
German	0.33	0.35	0.37	0.35	0.40	0.41	0.45	0.45	**0.46**
Italian	0.33	0.38	0.37	0.31	0.37	0.39	0.49	**0.52**	**0.52**
Portuguese	0.32	0.38	0.31	0.28	0.43	**0.46**	0.44	0.44	0.43
Spanish	0.38	0.39	0.39	0.35	0.42	0.41	**0.50**	0.49	**0.50**
Swedish	0.46	0.42	0.36	0.36	0.41	0.44	**0.49**	0.46	0.46
Avg.	0.36	0.38	0.38	0.35	0.40	0.42	**0.48**	0.47	**0.48**

Table 3: Micro-Averaged F-measure for sentence classification. Best scores are given in bold.

Approach	r	ρ
BoC	0.18	0.20
HCNN	0.24	0.26
HNN	0.28	0.32
Joint$_{doc}$	0.30	0.37
Joint	0.46	0.54
Joint$_{struc}$	**0.50**	**0.63**

Table 4: RILE score prediction performance. Best scores are given in bold.

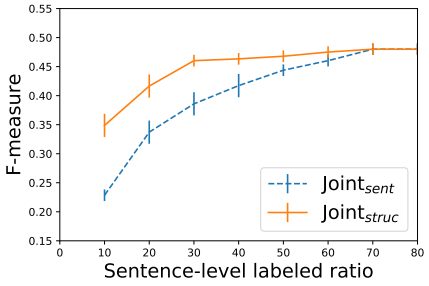


Figure 3: F-measure for *Joint$_{struc}$* vs. *Joint$_{sent}$* across different ratios of sentence-level labeled manifestos (averaged over 10 runs, with standard deviation)

ing. From Figure 3, having more manifestos with document-level supervision demonstrates the advantage of semi-supervised learning, especially when the sentence-level supervision is sparse ($\leq$ 40%)— *Joint$_{struc}$* performs better than *Joint$_{sent}$*.

5.5 Manifesto Position Re-ranking using PSL

Finally, we present the results using PSL, which calibrates the overall manifesto position on the left–right spectrum, obtained using the joint-structured model (*Joint$_{struc}$*). As we evaluate the effect of temporal dependency, we use manifestos before 2008-09 for training (868 in total) and the later ones (until 2015, 136 in total) for testing. This test set covers one recent set of election manifestos for most countries, and two for the Netherlands, Spain and United Kingdom. To avoid variance in right-to-left ratio and the target variable (`pos`, initialized using *Joint$_{struc}$*) between the training and test sets, we build a stacked network (Fast and Jensen, 2008), whereby we estimate values for the training set using cross-validation across the training partition, and estimate values for the test-set with a model trained over the entire training data. Note that we build the *Joint$_{struc}$* model afresh using the chronologically split training set, and the parameters are tuned again using an 80-20 random split of the training set. For a consistent view of results for both the tasks (and stages), we provide micro-averaged results for sentence-classification with the competing approaches (from Table 3): *AE-NN* (Subramanian et al., 2017), *Bi-LSTM(+up)*, and *Joint$_{struc}$*. Results are presented in Table 5, noting that the results for a given method will differ from earlier due to the different data split.

Approach	F-measure
AE-NN	0.31
Bi-LSTM(+up)	0.36
Joint$_{struc}$	**0.42**

Table 5: Micro-averaged F-measure for manifestos released after 2008-09. Best scores are given in bold.

For the document-level regression task, we also evaluate other approaches based on manifesto similarity and automated scaling with sentence-level policy positions:

- **Cross-lingual scaling (*CLS*)**: A recent unsupervised approach for crosslingual political speech text scoring (Glavaš et al., 2017b), based on TF-IDF weighed average word-embeddings to represent documents, and a graph constructed using pair-wise document

	RILE		CHES	
	r	ρ	r	ρ
CLS	0.11	0.10	0.09	0.07
PCA	0.26	0.17	0.01	−0.02
$Joint_{struc}$	0.46	0.42	0.42	0.42
PSL_{coal}	0.51	0.45	0.49	0.45
$PSL_{coal + esim}$	0.52	0.47	0.50	0.46
$PSL_{coal + esim + ploc}$	**0.54**	0.56	0.53	0.56
$PSL_{coal + esim + ploc + temp}$	**0.54**	**0.57**	**0.55**	**0.61**

Table 6: Manifesto regression task using the two-stage approach. Best scores are given in bold.

similarity. Given two pivot texts (for left and right), label propagation approach is used to position other documents.

- **PCA**: Apply principal component analysis (Gabel and Huber, 2000) on the distribution of sentence-level policy positions (56 classes, without 000), and use the projection on its principal component to explain maximum variance in its sentence-level positions, as a latent manifesto-level position score.

- ***Joint*$_{struc}$**: We evaluate the scores obtained using $Joint_{struc}$, which we calibrate using PSL.

We validate the calibrated position scores using both RILE and CHES[7] scores. We use CHES 2010-14, and map the manifestos to the closest survey year (wrt its election date). CHES scores are used only for evaluation and not during training. We provide results in Table 6 by augmenting features for the PSL model (Table 1) incrementally. We observed that the coalition-based feature, and polarity of sentences with its position information improves the overall ranking (r, ρ). Document similarity based relational feature provides only mild improvement (similarly to Burford et al. (2015)), and temporal dependency provides further improvement against CHES. That is, combining content, network and temporal features provides the best results.

6 Conclusion and Future Work

This work has been targeted at both fine- and coarse-grained manifesto text position analysis. We have proposed a two-stage approach, where in the first step we use a hierarchical multi-task deep model to handle the sentence- and document-level tasks together. We also utilize additional information on label structure, to augment an auxiliary structured loss. Since the first step places the manifesto on the left–right spectrum using text only, we leverage context information, such as coalition and temporal dependencies to calibrate the position further using PSL. We observed that: (a) a hierarchical bi-LSTM model performs best for the sentence-level classification task, offering a 10% improvement over the state-of-art approach (Subramanian et al., 2017); (b) modeling the document-level task jointly, and also augmenting the structured loss, gives the best performance for the document-level task and also helps the sentence-level task under sparse supervision scenarios; and (c) the inclusion of a calibration step with PSL provides significant gains in performance against both RILE and CHES, in the form of an increase from $\rho = 0.42$ to 0.61 wrt CHES survey scores.

There are many possible extensions to this work, including: (a) learning multilingual word embeddings with domain information; and (b) modeling other policy related scores from text, such as "support for EU integration".

[7] https://www.chesdata.eu/

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. This work was funded in part by the Australian Government Research Training Program Scholarship, and the Australian Research Council.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. EMNLP, pages 2289–2294.

S. H. Bach, B. Huang M. Broecheler, and L. Getoor. 2015. Hinge-loss markov random fields and probabilistic soft logic. *CoRR* abs/1505.04406.

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*.

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.

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

Felix Biessmann. 2016. Automating political bias prediction. *CoRR* abs/1608.02195.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

B. Bruinsma and K. Gemenis. 2017. Validating Wordscores. *CoRR* abs/1707.04737.

I Budge, H.D Klingemann, A Volkens, J Bara, E Tannenbaum, R Fording, D Hearl, H.M Kim, M McDonald, and S Mendes. 2001. *Mapping Policy Preferences: Parties, Electors and Governments*. Oxford University Press.

Clint Burford, Steven Bird, and Timothy Baldwin. 2015. Collective document classification with implicit inter-document semantic relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*. Denver, USA, pages 106–116.

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` .

Thomas Däubler, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2012. Natural sentences as valid units for coded political texts. *British Journal of Political Science* 42(4):937–951.

Andrew Fast and David Jensen. 2008. Why stacked models perform effective collective classification. In *Proceedings of the Eighth International Conference on Data Mining*. IEEE, pages 785–790.

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.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017a. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. ACL, pages 42–46.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017b. Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. volume 2, pages 688–693.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, pages 6645–6649.

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.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21(3):267–297.

Frederik Georg Hjorth, Robert Tranekær Klemmensen, Sara Binzer Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. 2015. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research and Politics* 2(2):1–9.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter. In *Proceedings of the Association for Computational Linguistics*. ACL, pages 741–752.

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

Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping Policy Preferences II. Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD*. Oxford University Press.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323(5915):721.

Rémi Lebret and Ronan Collobert. 2014. N-gram-based low-dimensional representation for document classification. *CoRR* abs/1412.6277.

Jiwei Li, Luong Minh-Thang, and Jurafsky Dan. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. ACL, pages 1106–1115.

James Lo, Sven-Oliver Proksch, and Thomas Gschwend. 2013. A common left-right scale for voters and parties in europe. *Political Analysis* 22(2):205–223.

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

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. ACL, pages 432–439.

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in us electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. EMNLP, pages 2938–2944.

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).

Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis* 20(1):78–91.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)* 17(3):26.

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.

Samuel L Smith, Turban David HP, Hamblin Steven, and Hammerla Nils Y. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*. ACL, pages 109–117.

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin, and Julian Brooke. 2017. Joint sentence-document model for manifesto text analysis. In *Proceedings of the 15th Annual Workshop of The Australasian Language Technology Association*. ALTA, pages 25–33.

Suzan Verberne, Eva Dhondt, 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*, Oxford University Press, chapter 6.

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

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*. PolText, pages 88–93.

3.2 Campaign text speech act analysis [*SEM 2019]

Our earlier work (Section 3.1), focused on sentence- and document-level tasks for manifesto text analysis, namely policy theme classification and ideology regression. It is known that election speeches and media-releases contain policy related functions and dynamic interactions between parties, which are used by each party to place itself ahead in the electoral contest [Gautrais et al. (2017)]. Unlike the work with the CMP dataset, we do not have any standard label schema or political speech dataset annotated for speech act analysis. Hence for political discourse analysis, we propose a speech act classification task by combining traditional speech act types with other speech act types tailored to political analysis [Eder et al. (2017)]. The key idea is to analyze the context of text in election campaigns, i.e., if a sentence discusses *military*, we want to know whether it is about the party's *past-action*, or a *judgement* about opposite party's policies or a *promise* on its own future action. This context can help to understand the party position more clearly, and capture issue salience in a more nuanced way, thereby characterizing parties better. This task contains two sub-tasks: (a) to distinguish between actions of the party making the utterance and actions of other parties, i.e., the *target* of each utterance has to be classified; and (b) since a sentence can have multiple types of speech acts and/or *target* party, we require sentence segmentation as a preliminary step. For this task definition, we developed a dataset using 2016 Australian election texts — speech transcripts and media releases, from the two major parties, Labor and Liberal. This corpus contains speeches in English text only, and is not multi-lingual as in our previous work with manifestos.

During the development of this work, language models were getting popular and showed promising results compared to bag-of-words style trained word embeddings. Hence in this work we use deep contextualized word representation in the form of ELMo [Peters et al. (2018)], to initialize word embeddings. We use a bidirectional Gated Recurrent Unit model for both speech acts and target party classification tasks. Since it is difficult to get access to large amounts of training data, we also proposed a semi-supervised extension of the model using a teacher-student framework, where the student model shares parameters with the teacher (main) network but is exposed to a restricted view of the input, and the framework enforces consensus between the two models. For segmenting sentence into basic utterances, such that each utterance exhibits a single action and owned by a single *target* party, we use a Conditional Random Field model. We address the segmentation and classification tasks in a cascaded fashion. We observe that the sequential model, and its semi-supervised extension performs best. We performed exploratory analysis to

identify differences in rhetorical function usage across policy issues, between the two parties (Labor and Liberal) on the opposite sides of ideological spectrum. Other than the exploratory analysis, predicting the party ideology as a function of speech act type distribution can help to analyze the superiority of this fine-grained representation quantitatively over just using policy mentions (as done in Section 3.1). This is not feasible here because we are dealing with text from a single election cycle, which leaves us with just one score per party, hence the prediction task is open for future exploration covering more data from many parties across multiple elections. This work was published at *SEM 2019, co-located with NAACL-HLT 2019.

Target Based Speech Act Classification in Political Campaign Text

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

We study pragmatics in political campaign text, through analysis of speech acts and the target of each utterance. We propose a new annotation schema incorporating domain-specific speech acts, such as *commissive-action*, and present a novel annotated corpus of media releases and speech transcripts from the 2016 Australian election cycle. We show how speech acts and target referents can be modeled as sequential classification, and evaluate several techniques, exploiting contextualized word representations, semi-supervised learning, task dependencies and speaker meta-data.

1 Introduction

Election campaign text is a core artifact in political analysis. Campaign communication can influence a party's reputation, credibility, and competence, which are primary factors in voter decision making (Fernandez-Vazquez, 2014). Also, modeling the discourse is key to measuring the role of party in *constructive democracy* — to engage in constructive discussion with other parties in a democracy (Gibbons et al., 2017).

Speech act theory (Austin, 1962; Searle, 1976) can be used to study such pragmatics in political campaign text. Traditional speech act classes have been studied to analyze how people engage with elected members (Hemphill and Roback, 2014), and how elected members engage in discussions (Shapiro et al., 2018), with a particular focus on pledges (Artés, 2011; Naurin, 2011, 2014; Gibbons et al., 2017). Also, election manifestos have been analyzed for prospective and retrospective messages (Müller, 2018). In this work, we combine traditional speech acts with those proposed by political scientists to study political discourse, such as *specific pledges*, which can also help to verify the pledges' fulfilment after an election (Thomson et al., 2010).

In addition to speech acts, it is important to identify the target of each utterance — that is, the political party referred to in the text — in order to determine the discourse structure. Here, we study the effect of jointly modeling the speech act and target referent of each utterance, in order to exploit the task dependencies. That is, this paper is an application of discourse analysis to the pragmatics-rich domain of political science, to determine the intent of every utterance made by politicians, and in part, automatically extract pledges at varying levels of specificity from campaign speeches and press releases.

We assume that each utterance is associated with a unique speech act (similar to Zhao and Kawahara (2017)) and target party,[1] meaning that a sentence with multiple speech acts and/or targets must be segmented into component utterances. Take the following example, from the Labor Party:

(1) Labor will contribute $43 million towards the Roe Highway project and we call on the WA Government to contribute funds to get the project underway.

The example is made up of two utterances (with and without an underline), belonging to speech act types *commissive-action-specific* and *directive*, referring to different parties (LABOR and LIBERAL), resp. In our initial experiments, we perform target based speech act classification (i.e. joint speech act classification and determination of the target of the utterance) over gold-standard utterance data (Section 6), but return to perform automatic utterance segmentation along with target based speech act classification (Section 7).

While speech act classification has been applied to a wide range of domains, its application to polit-

[1] Zhao and Kawahara (2017) do not address the target referent classification task in their work.

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 273–282
Minneapolis, June 6–7, 2019.

Utterance	Speech act	Target party	Speaker
Tourism directly and indirectly supports around 38000 jobs in TAS.	*assertive*	NONE	LABOR
We will invest $25.4 million to increase forensics and intelligence assets for the Australian Federal Police	*commissive-action-specific*	LIBERAL	LIBERAL
Labor will prioritise the Metro West project if elected to government.	*commissive-action-vague*	LABOR	LABOR
A Shorten Labor Government will create 2000 jobs in Adelaide.	*commissive-outcome*	LABOR	LABOR
Federal Labor today calls on the State Government to commit the final $75 million to make this project happen.	*directive*	LIBERAL	LABOR
Good morning everybody.	*expressive*	NONE	LABOR
The Coalition has already delivered a $2.5 billion boost to our law enforcement and security agencies.	*past-action*	LIBERAL	LIBERAL
Malcolm Turnbull's health cuts will rip up to $1.4 billion out of Australians' pockets every year	*verdictive*	LIBERAL	LABOR

Table 1: Examples with speech act and target party classes. "Speaker" denotes the party making the utterance.

ical text is relatively new. Most speech act analyses in the political domain have relied exclusively on manual annotation, and no labeled data has been made available for training classifiers. 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 utilizing unlabeled text, which is easy to obtain.

The contributions of this paper are as follows: (1) we introduce the novel task of target based speech act classification to the analysis of political discourse; (2) we develop and release a dataset (can be found here `https://github.com/shivashankarrs/Speech-Acts`) based on political speeches and press releases, from the two major parties — Labor and Liberal — in the 2016 Australian federal election cycle; and (3) we propose a semi-supervised learning approach to the problem by augmenting the training data with in-domain unlabeled text.

2 Related Work

The recent adoption of NLP methods has led to significant advances in the field of computational social science (Lazer et al., 2009), including political science (Grimmer and Stewart, 2013). With the increasing availability of datasets and computational resources, large-scale comparative political text analysis has gained the attention of political scientists (Lucas et al., 2015). One task of particular importance is the analysis of the functional intent of utterances in political text. Though it has received notable attention from many political scientists (see Section 1), the primary focus of almost all work has been to derive insights from manual annotations, and not to study computational approaches to automate the task.

Another related task in the political communication domain is reputation defense, in terms of party credibility. Recently, Duthie and Budzynska (2018) proposed an approach to mine ethos support/attack statements from UK parliamentary debates, while Naderi and Hirst (2018) focused on classifying sentences from Question Time in the Canadian parliament as defensive or not. In this work, our source data is speeches and press releases in the lead-up to a federal election, where we expect there to be rich discourse and interplay between political parties.

Speech act theory is fundamental to study such discourse and pragmatics (Austin, 1962; Searle, 1976). A speech act is an illocutionary act of conversation and reflects shallow discourse structures of language. Due to its predominantly small-data setting, speech act classification approaches have generally relied on bag-of-words models (Qadir and Riloff, 2011; Vosoughi and Roy, 2016), although recent approaches have used deep-learning models through data augmentation (Joty and Hoque, 2016) and learning word representations for the target domain (Joty and Mohiuddin, 2018), outperforming traditional bag-of-words approaches.

Another technique that has been applied to com-

pensate for the sparsity of labeled data is semi-supervised learning, making use of auxiliary unlabeled data, as done previously for speech act classification in e-mail and forum text (Jeong et al., 2009). Zhang et al. (2012) also used semi-supervised methods for speech act classification over Twitter data. They used transductive SVM and graph-based label propagation approaches to annotate unlabeled data using a small seed training set. Joty and Mohiuddin (2018) leveraged out-of-domain labeled data based on a domain adversarial learning approach. In this work, we focus on target based speech act analysis (with a custom class-set) for political campaign text and use a deep-learning approach by incorporating contextualized word representations (Peters et al., 2018) and a cross-view training framework (Clark et al., 2018) to leverage in-domain unlabeled text.

3 Problem Statement

Target based speech act classification requires the segmentation of sentences into utterances, and labelling of those utterances according to speech act and target party. In this work we focus primarily on speech act and target party classification.

Our speech act coding schema is comprised of: *assertive*, *commissive*, *directive*, *expressive*, *past-action*, and *verdictive*. An *assertive* commits the speaker to something being the case. With a *commissive*, the speaker commits to a future course of action. Following the work of Artés (2011) and Naurin (2011), we distinguish between *action* and *outcome* commissives. Action commissives (*commissive-action*) are those in which an action is to be taken, while outcome commissives (*commissive-outcome*) can be defined as a description of reality or goals. Secondly, similar to Naurin (2014) we also classify action commissives into vague (*commissive-action-vague*) and specific (*commissive-action-specific*), according to their specificity. This distinction is also related to text specificity analysis work addressed in the news (Louis and Nenkova, 2011) and classroom discussion (Lugini and Litman, 2017) domains. A *directive* occurs when the speaker expects the listener to take action in response. In an *expressive*, the speaker expresses their psychological state, while a *past-action* denotes a retrospective action of the target party, and a *verdictive* refers to an assessment on prospective or retrospective actions.

Examples of the eight speech act classes are given in Table 1, along with the target party (LABOR, LIBERAL, or NONE), indicating which party the speech act is directed towards, and the "speaker" party making the utterance (information which is provided for every utterance).

# Doc	# Sent	# Utt	Avg Utterance Length
258	6609	7641	19.3

Table 2: Dataset Statistics: number of documents, number of sentences, number of utterances, and average utterance length

3.1 Utterance Segmentation

Sentences are segmented both in the context of speech act and target party — when a sentence has utterances belonging to more than one speech act or/and more than one target. For example, the following sentence conveys a pledge (*commissive-outcome*) followed by the party's belief (*assertive*), with the utterance boundary indicated by |:

(2) We will save Medicare | because Medicare is more than just a standard of health.

Further, the following (from the Labor party) has segments comparing LABOR and LIBERAL:

(3) Our party is united – | the Liberals are not united.

4 Election Campaign Dataset

We collected media releases and speeches from the two major Australia political parties — Labor and Liberal — from the 2016 Australian federal election campaign. A statistical breakdown of the dataset is given in Table 2. We compute agreement over 15 documents, annotated by two independent annotators, with disagreements resolved by a third annotator. The remaining documents are annotated by the two main annotators without redundancy. Agreement between the two annotators for utterance segmentation based on exact boundary match using Krippendorff's alpha (α) (Krippendorff, 2011) is 0.84. Agreement statistics for the classification tasks (Cohen, 1960; Carletta, 1996) are given in Tables 3 and 4.

Speech act	%	Kappa (κ)
assertive	40.8	0.85
commissive-action-specific	12.4	0.84
commissive-action-vague	6.6	0.73
commissive-outcome	4.9	0.72
directive	1.7	0.92
expressive	1.9	0.88
past-action	6.3	0.76
verdictive	25.4	0.82

Table 3: Speech act agreement statistics

Target party	%	Kappa (κ)
LABOR	45.9	0.92
LIBERAL	39.1	0.90
NONE	15.0	0.86

Table 4: Target party agreement statistics

5 Proposed Approach

Our approach to labeling utterances for speech act and target party classification is as follows. Utterances are first represented as a sequence of word embeddings, and then using a bidirectional Gated Recurrent Unit ("biGRU": Cho et al. (2014)). The representation of each utterance is set to the concatenation of the last hidden state of both the forward and backward GRUs, $\mathbf{h}_i = \left[\overrightarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{h}}_i\right]$. After this, the model has a softmax output layer. This network is trained for both the speech act (eight class) and target party (three class) classification tasks, minimizing cross-entropy loss, denoted as $\mathcal{L}_S$ and $\mathcal{L}_T$ respectively.

Our approach has the following components:

ELMo word embeddings ("biGRU$_{\text{ELMo}}$"): As word embeddings we use a 1024d learned linear combination of the internal states of a bidirectional language model (Peters et al., 2018).

Semi-supervised Learning: We employ a cross-view training approach (Clark et al., 2018) to leverage a larger volume of unlabeled text. Cross-view training is a kind of teacher–student method, whereby the model "teaches" another "student" model to classify unlabelled data. The student sees only a limited form of the data, e.g., through application of noise (Sajjadi et al., 2016; Wei et al., 2018), or a different view of the input, as used herein. This procedure regularises the learning of the teacher to be more robust, as well as increasing the exposure to unlabeled text.

We augment our dataset with over 36k sentences from Australian Prime Minister candidates' election speeches.[2] On these unlabeled examples, the model's probability distribution over targets $p_\theta(y|s)$ is used to fit auxiliary model(s), $p_\omega(y|s)$, by minimising the Kullback-Leibler (KL) divergence, $\text{KL}(p_\theta(y|s), p_\omega(y|s))$. This consensus loss component, denoted $\mathcal{L}_{\text{unsup}}$, is added to the supervised training objective ($\mathcal{L}_S$ or $\mathcal{L}_T$).

We evaluate the following auxiliary models:[3]

- a forward GRU ("biGRU$_{\text{CVT}_{\text{fwd}}}$");
- separate forward and backward GRUs ("biGRU$_{\text{CVT}_{\text{fwdbwd}}}$"); and
- a biGRU with word-level dropout ("biGRU$_{\text{CVT}_{\text{worddrop}}}$").

The intuition is that the student models only have access to restricted views of the data on which the teacher network is trained, and therefore this acts as a regularization factor over the unlabeled data when learning the teacher model.

Multi-task Learning ("biGRU$_{\text{Multi}}$"): For speech act classification, target party classification is considered as an auxiliary task, and vice versa. Accordingly, a separate model is built for each task, with the other task as an auxiliary task, in each case using a linearly weighted objective $\mathcal{L}_S + \alpha\mathcal{L}_T$, where $\alpha \geq 0$ is tuned separately in each application. The intuition here is to capture the dependencies between the tasks, e.g., *commissive* is relevant to the Speaker party only.

Meta-data (biGRU$_{\text{Meta}}$): We concatenate a binary flag encoding the speaker party ($\mathbf{m}_i$) alongside the utterance embedding $\mathbf{h}_i$, i.e., $[\mathbf{h}_i, \mathbf{m}_i]$. This representation is passed through a hidden layer with ReLU-activation, then projected onto a output layer with softmax activation for both the classification tasks.

6 Evaluation

We compare the models presented in Section 5 with the following baseline approaches:

- Support Vector Machine ("SVM$_{\text{BoW}}$") with with unigram term-frequency representation.
- Multi-layer perceptron ("MLP$_{\text{BoW}}$") with unigram term-frequency representation.

[2] https://primeministers.moadoph.gov.au/collections/election-speeches

[3] Note that auxliary models share parameters with the corresponding components of main (teacher) model, with the exception of their output layers.

ID	Approach	Speech act		Target party	
		Accuracy	Macro-F1	Accuracy	Macro-F1
1	$\text{Meta}_{\textbf{naive}}$	—	—	0.55	0.43
2	$\text{SVM}_{\textbf{BoW}}$	0.56	0.41	0.60*[1]	0.56*[1]
3	$\text{MLP}_{\textbf{BoW}}$	0.60*[2]	0.47*[2]	0.61*[1]	0.57*[1]
4	$\text{DAN}_{\textbf{GloVe}}$	0.53	0.30	0.59	0.54
5	$\text{GRU}_{\textbf{GloVe}}$	0.56	0.46	0.58	0.55
6	$\text{biGRU}_{\textbf{GloVe}}$	0.57	0.48	0.59	0.56
7	$\text{MLP}_{\textbf{ELMo}}$	0.62*[3]	0.53*[3]	0.58	0.57
8	$\text{biGRU}_{\textbf{ELMo}}$	**0.68***[7]	**0.57***[7]	0.63*[2,3,7]	0.60*[2,3,7]
9	$\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{fwd}}}$	0.66	0.55	0.63	0.58
10	$\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{fwdbwd}}}$	**0.68**	0.54	0.61	0.56
11	$\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{worddrop}}}$	**0.69**	**0.57**	0.66*[8]	0.60
12	$\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{worddrop}}\textbf{ + Multi}}$	**0.69**	**0.58**	0.65	0.60
13	$\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{worddrop}}\textbf{ + Meta}}$	**0.68**	**0.58**	**0.71***[11]	**0.66***[8,11]

Table 5: Classification results showing average performance based on 10 runs. * indicates results significantly better than the indicated approaches (based on ID in the table) according to a paired t-test ($p < 0.05$). Boldface shows the overall best results and results insignificantly different from the best. $\text{Meta}_{\textbf{naive}}$ is not applicable for speech act classification. Note that all approaches use gold-standard segmentation for evaluation.

Speech act	MLP_{ELMo}	Our approach
assertive	0.77	0.80
commissive-action-specific	0.65	0.69
commissive-action-vague	0.45	0.48
commissive-outcome	0.28	0.39
directive	0.58	0.59
expressive	0.55	0.58
past-action	0.45	0.48
verdictive	0.48	0.61

Table 6: Speech act class-wise F1 score.

Target party	$\text{biGRU}_{\text{ELMo}}$	Our approach
LABOR	0.68	0.74
LIBERAL	0.65	0.75
NONE	0.46	0.48

Table 7: Target party class-wise F1 score.

- Deep Averaging Networks ("$\text{DAN}_{\textbf{GloVe}}$") (Iyyer et al., 2015), GRU ("$\text{GRU}_{\textbf{GloVe}}$"), and biGRU ("$\text{biGRU}_{\textbf{GloVe}}$") with pre-trained 300d GloVe embeddings (Pennington et al., 2014).
- MLP with average-pooling over pre-trained ELMo word embeddings ("$\text{MLP}_{\textbf{ELMo}}$").
- Using speaker party as the predicted target party ("$\text{Meta}_{\textbf{naive}}$").

We average results across 10 runs with 90%/10% training/test random splits. Hyperparameters are tuned over a 10% validation set randomly sampled and held out from the training set. We evaluate using accuracy and macro-averaged F-score, to account for class-imbalance. We compare the baseline approaches against our proposed approach (different components given in Section 5). We evaluate the effect of each component by adding them to the base model ($\text{biGRU}_{\textbf{ELMo}}$), e.g., biGRU model with ELMo embeddings and word-level dropout based semi-supervised approach is given as $\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{worddrop}}}$. Results for speech act and target party classification are given in Table 5. The corresponding class-wise performance for both speech act and target party tasks with our approach ($\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{worddrop}}\textbf{ + Meta}}$) compared against the competitive approach from Table 5 is given in Table 6 and Table 7 respectively (and also discussed further in Section 8). All the approaches are evaluated with the gold-standard segmentation. Utterance segmentation is discussed in Section 7.

From the results in Table 5, we observe that the biGRU[4] performs better than the other approaches, and that ELMo contextual embeddings ($\text{biGRU}_{\textbf{ELMo}}$) boosts the performance appreciably. Apart from ELMo, the semi-supervised learning methods ($\text{biGRU}_{\textbf{ELMo + CVT}_{\textbf{worddrop}}}$) provide a boost in performance for the target party

[4]The biGRU model uses ReLU activations, a 128d hidden layer for speech act classification and 64d hidden layer for target party classification, and dropout rate of 0.1.

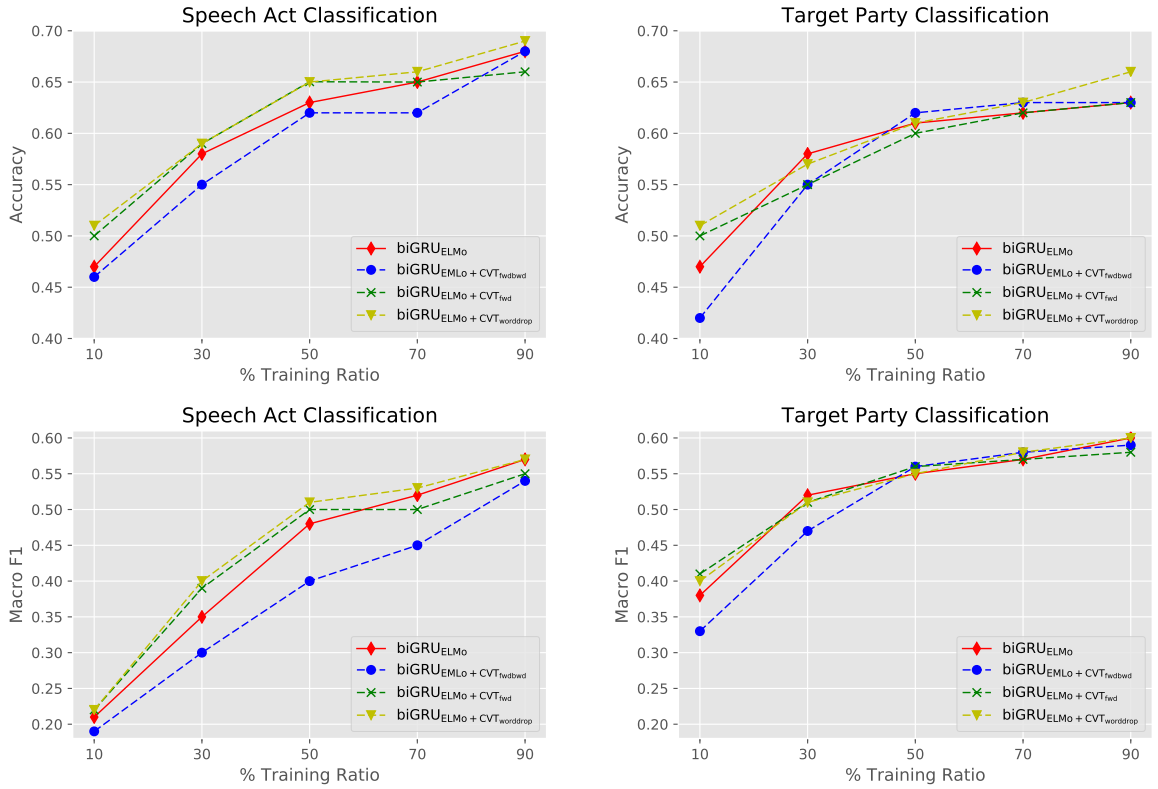


Figure 1: Classification performance across different training ratios. Note that 90% is using all the training data, as 10% is used for validation.

task (wrt accuracy) using all the training data. $\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}}$ and $\text{biGRU}_{\text{ELMo + CVT}_{\text{fwd}}}$ provide gains in performance for the speech act task, especially with fewer training examples ($\leq$ 50% of training data, see Figure 1). Performance of semi-supervised learning models with cross-view training (which leverages in-domain unlabeled text) is compared against $\text{biGRU}_{\text{ELMo}}$, which is a supervised approach. Results across different training ratio settings are given in Figure 1. From this, we can see that $\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}}$ and $\text{biGRU}_{\text{ELMo + CVT}_{\text{fwd}}}$ performs better than $\text{biGRU}_{\text{ELMo + CVT}_{\text{fwdbwd}}}$ in almost all cases. With a training ratio $\leq$ 50%, $\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}}$ achieves a comparable performance to $\text{biGRU}_{\text{ELMo + CVT}_{\text{fwd}}}$.

Multi-task learning ($\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}\text{ + Multi}}$) provides only small improvements for the speech act task. Further, when we add speaker party meta-data ($\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}\text{ + Meta}}$), it provides large gains in performance for the target party task. Overall, the proposed approach ($\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}\text{ + Meta}}$) provides the best performance for the target party task. Its performance is better than the $\text{biGRU}_{\text{ELMo + Meta}}$ model, which does not leverage the additional unlabeled text using semi-supervised learning, where it achieves 0.70 accuracy and 0.65 Macro F1. Also, ELMo and semi-supervised methods ($\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}}}$ and $\text{biGRU}_{\text{ELMo + CVT}_{\text{fwd}}}$) provide significant improvements for the speech act task, especially under sparse supervision scenarios (see Figure 1, for training ratio $\leq$ 50%).

7 Segmentation Results

In the previous experiments, we used gold-standard utterance data, but next we experiment with automatic segmentation. We use sentences as input, based on the NLTK sentence tokenizer (Bird et al., 2009), and automatically segment sentences into utterances based on token-level segmentation, in the form of a `BI` binary sequence classification task using a CRF model (Hernault et al., 2010).[5] We use the following set of features for each word: token, word shape (capitalization, punctuation, digits), Penn POS tags based on SpaCy, ClearNLP dependency labels (Choi and Palmer, 2012), relative position in the sentence, and features for the

[5]We also experimented with a neural CRF model, but found it to be less accurate.

Utterance	Target party	Speaker
Our new Tourism Infrastructure Fund will bring more visitor dollars and more hospitality jobs to Cairns, Townsville and the regions.	LABOR	LABOR
Just as he sold out 35,000 owner-drivers in his deal with the TWU to bring back the “Road Safety Remuneration Tribunal".	LABOR	LIBERAL
Then in 2022, we will start construction of the first of 12 regionally superior submarines, the single biggest investment in our military history.	LIBERAL	LIBERAL

Table 8: Scenarios where “Speaker” meta-data benefits the target party classification task.

adjacent words (based on this same feature representation). We compute segmentation accuracy (SA: Zimmermann et al. (2006)), which measures the percentage of segments that are correctly segmented, i.e. both the left and right boundary match the reference boundaries. SA for the CRF model is 0.87. Secondly, to evaluate the effect of segmentation on classification, we compute joint accuracy (JA). It is similar to SA but also requires correctness of the speech act and target party. In cascaded style, JA using the CRF model for segmentation and $\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}} + \text{Meta}}$ for speech act and target party classification is 0.60 and 0.64 respectively. Here, segmentation errors lead to a small drop in performance.

8 Error Analysis

We analyze the class-wise performance and confusion matrix for our best performing approach ($\text{biGRU}_{\text{ELMo + CVT}_{\text{worddrop}} + \text{Meta}}$). Speech act and target party class-wise performance is given in Tables 6 and 7 respectively. We can see that the proposed approach provides improvement across all classes, while achieving comparable performance for *directive*. Recognizing *commissive-outcome* can be seen to be tougher than other classes. In addition, we analyze the results to identify cases where having “Speaker” party information is beneficial for predicting the target party of sentences. Some of those scenarios are given in Table 8, where the meta-data enables predicting the target party correctly even when there is no explicit reference to the party or leaders.

Confusion matrices for the speech act and target party classification tasks are given in Figure 2. Some observations from the confusion matrices are: (a) *assertive* and *verdictive* are often misclassified as each other; (b) *commissive-action-vague* utterances are often misclassified as *commissive-action-specific*; and (c) LABOR and LIBERAL classes are often misclassified as each other for the target party classification task.

9 Qualitative Analysis

Here we provide the policy-wise speech act distribution for both parties, which indicates the difference in their predilection for the indicated six policy areas (Figure 3). We provide results for the six most frequent policy categories, for each of which, the campaign text is first classified into one of the policy-areas that are relevant to Australian politics, by building a Logistic Regression classifier with data obtained from ABC Fact Check.[6] Some observations (based on Figure 3) are as follows:

- The incumbent government (LIBERAL) uses more *directive*, *expressive*, *verdictive*, and *past-action* utterances than the opposition (LABOR).
- LIBERAL’s text has relatively more pledges (*commissive-action-vague*, *commissive-action-specific* and *commissive-outcome*) on **economy** compared to LABOR, whereas LABOR has more pledges on **social services** and **education**. This is as expected for right- and left-wing parties respectively. Other policy-areas have a comparable number of pledges from both parties. Overall, party-wise salience towards these policy areas correlates highly with the relative breakdowns in the Comparative Manifesto Project (Volkens et al., 2017): where the relative share of sentences from the LABOR and LIBERAL manifestos[7] for **welfare state (health and social services)** is 22:7, **education** is 9:6, **economy** is 11:23, and **technology & infrastructure (communication, infrastructure)** is 17:19.
- Across policy-areas, *specific* pledges are

[6] https://www.abc.net.au/news/factcheck

[7] https://manifesto-project.wzb.eu/down/data/2018b/datasets/MPDataset_MPDS2018b.csv

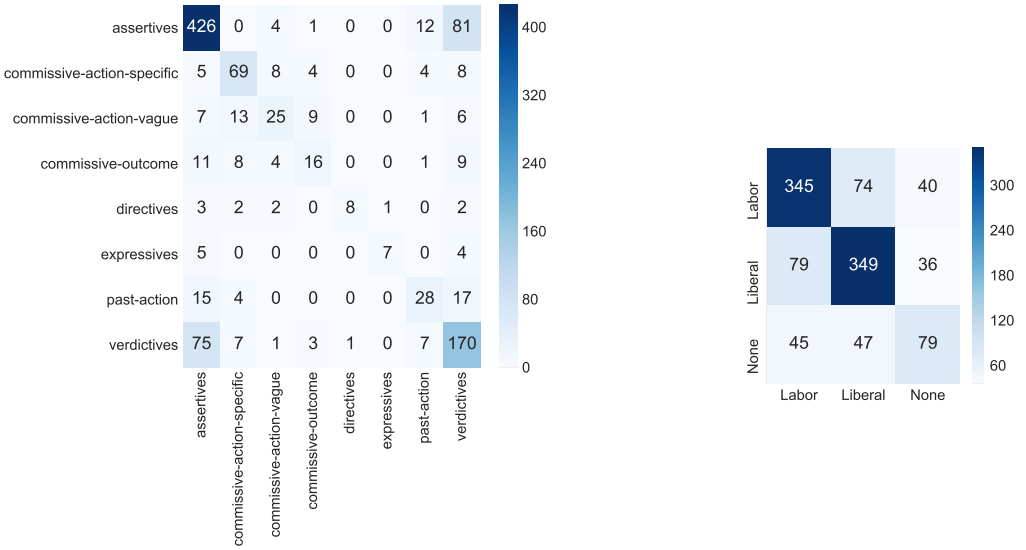


Figure 2: Confusion matrix for speech act and target party classification tasks.

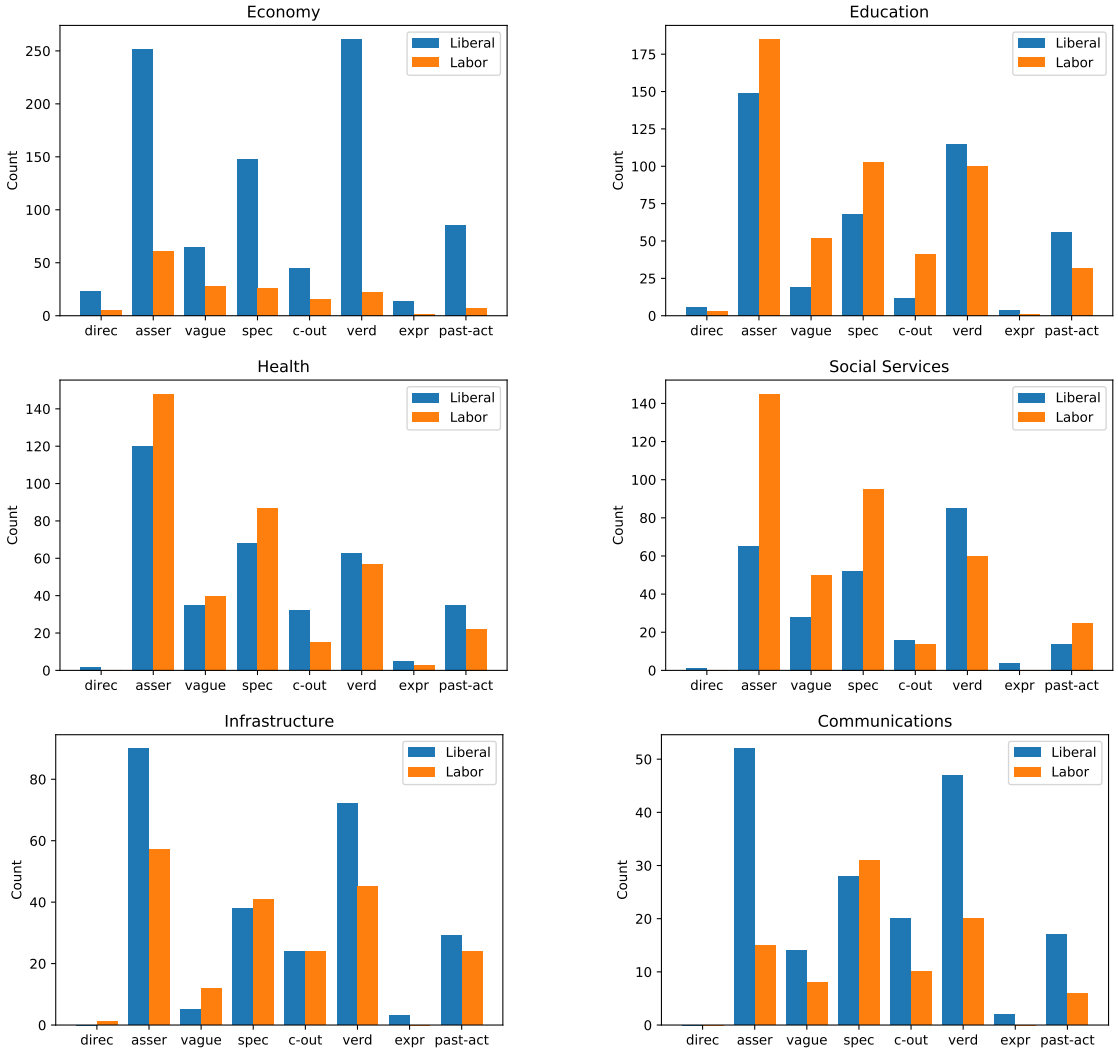


Figure 3: Policy-wise speech act analysis. Classes include: *directive* (“direc”), *assertive* (“asser”), *commissive-action-vague* (“vague”), *commissive-action-specific* (“spec”), *commissive-outcome* (“c-out”), *verdictive* (“verd”), *expressive* (“expr”), and *past-action* (“past-act”)

more frequent than *vague* ones. This aligns with previous studies done by Naurin (2014) and Gibbons et al. (2017).

10 Conclusion and Future Work

In this work we present a new dataset of election campaign texts, based on a class schema of speech acts specific to the political science domain. We study the associated problems of identifying the referent political party, and segmentation. We showed that this task is feasible to annotate, and present several models for automating the task. We use a pre-trained language model and also leverage auxiliary unlabeled text with semi-supervised learning approach for the target based speech act classification task. Our results are promising, with the best method being a semi-supervised biGRU with ELMo embeddings for the speech act task, and the model additionally incorporating speaker meta-data for the target party task. We provided qualitative analysis of speech acts across major policy areas, and in future work aim to expand this analysis further with fine-grained policies and ideology-related analysis.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. This work was funded in part by the Australian Government Research Training Program Scholarship, and the Australian Research Council.

References

J. Artés. 2011. Do Spanish politicians keep their election promises? *Party Politics*, 19(1):143–158.

J. L. Austin. 1962. *How to do things with words*. Clarendon Press, Oxford.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

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.

Jinho D Choi and Martha Palmer. 2012. Guidelines for the clear style constituent to dependency conversion. *Technical Report 01–12: Institute of Cognitive Science, University of Colorado Boulder*.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Rory Duthie and Katarzyna Budzynska. 2018. A deep modular RNN approach for ethos mining. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4041–4047.

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

Andrew Gibbons, Aaron Martin, and Andrea Carson. 2017. Do parties keep their election promises?: A study of the 43rd Australian Parliament (2010-2013). *Australian Political Studies Association*.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Libby Hemphill and Andrew J Roback. 2014. Tweet Acts: How Constituents Lobby Congress via Twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1200–1210.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A sequential model for discourse segmentation. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 315–326.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1681–1691.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259.

Shafiq Joty and Enamul Hoque. 2016. Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured

models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1746–1756.

Shafiq Joty and Tasnim Mohiuddin. 2018. Modeling speech acts in asynchronous conversations: A neural-CRF approach. *Computational Linguistics*, pages 1–36.

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

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: The coming age of computational social science. *Science*, 323(5915):721.

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.

Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2):254–277.

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.

Stefan Müller. 2018. Prospective and retrospective rhetoric: A new dimension of party competition and campaign strategies. In *Manifesto Corpus Conference*.

Nona Naderi and Graeme Hirst. 2018. Using context to identify the language of face-saving. In *Proceedings of the 5th Workshop on Argument Mining*, pages 111–120.

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

E. Naurin. 2014. Is a promise a promise? Election pledge fulfillment in comparative perspective using Sweden as an example. *West European Politics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

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.

Ashequl Qadir and Ellen Riloff. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 748–758.

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.

J. R. Searle. 1976. *A Taxonomy of Illocutionary Acts*. Linguistic Agency University of Trier.

Matthew A. Shapiro, Libby Hemphill, and Jahna Otterbacher. 2018. Updates to congressional speech acts on Twitter. In *Proceedings of the American Political Science Association Annual Meeting*.

Robert Thomson, Terry Royed, and Elin Naurin. 2010. The Program-to-Policy Linkage: A Comparative Study of Election Pledges and Government Policies in the United States, the United Kingdom, the Netherlands and Ireland. In *Proceedings of the American Political Science Association Annual Meeting*.

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.

Soroush Vosoughi and Deb Roy. 2016. Tweet acts: A speech act classifier for Twitter. In *Proceedings of the Tenth International AAAI Conference On Web And Social Media*, pages 711–715.

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*.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2012. Towards scalable speech act recognition in Twitter: tackling insufficient training data. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 18–27.

Tianyu Zhao and Tatsuya Kawahara. 2017. Joint learning of dialog act segmentation and recognition in spoken dialog using neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 704–712.

M Zimmermann, A Stolcke, and E Shriberg. 2006. Joint segmentation and classification of dialog acts in multiparty meetings. In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 581–584.

3.3 Election promises specificity prediction [EMNLP-IJCNLP 2019]

Our earlier work using election campaign text (Sections 3.1 and 3.2) dealt with the classification of policy themes and speech acts. In this work, we focus on a core part of elections: *promises*. We propose a new NLP task for promise specificity prediction based on a class schema provided by political scientists [Pomper and Lederman (1980)]. The key idea is to differentiate between promises such as *We will work to ensure Australia has a sustainable and fair retirement income system* and *We will lower the annual non-concessional contributions cap to $75,000 and lower the High Income Superannuation Contribution threshold to $200,000.* Vague promises (such as the first) convey less information and have less accountability than more specific ones (such as the second promise).

We model each sentence using a bi-GRU initialized with ELMo word embeddings, a model that worked well for speech act classification (Section 3.2). Rather than modeling the output as a regression or classification target, we model it as an ordinal regression task. The intuition behind this is that the output classes have a grade (they are not at the same level, because one is more specific than another), and the difference in commitment between two categories of promise specificity is not linear. Since creating large amounts of labeled data is hard, as in the case of any other domain specialized task, we propose a semi-supervised extension of the model using a teacher-student framework (similar to Section 3.2), wherein we use a custom consensus loss for the ordinal output. Finally, we quantitatively show that promises and their specificity can characterize party ideology and issue salience better than simple policy-theme mentions as done in Section 3.1, which is also a popularly used approach in existing literature. We model ideology as a function of promise specificity across policy issues using a PSL model, which allows us to incorporate contextual information such as party coalition and temporal dependencies (as done for the policy theme and ideology regression task, in Section 3.1) and we use a linear model for issue salience analysis. We developed a dataset using Australian manifestos over twelve election cycles for evaluation.

We observe that the sequential model using a bi-GRU with ELMo performs better than average embedding settings, and ordinal output modeling is better than both classification and regression models. In the semi-supervised extension, we find that consensus loss which captures the ordinal nature of the task performs best. Lastly, with the PSL and linear model, we found that specific promises made on different policy issues can better characterize parties than simple policy theme mentions. In this work, similar to Section 3.1, we have a two-stage approach for the lower-level

text prediction task and the higher-level party characteristic analysis. This work was published at EMNLP-IJCNLP 2019.

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

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1729–1740, Hong Kong, China, November 3–7, 2019.

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

[1] https://github.com/shivashankarrs/Pledge-Specificity

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.

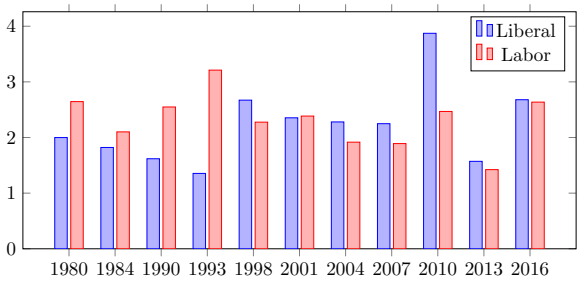


Figure 1: Average pledge specificity of Labor and Liberal parties over different election cycles

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 crowdsourcing 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 results 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, $q_x : \mathcal{Y} \rightarrow [0, 1]$. For this, the ordinal label of training instances is transformed into a truncated Gaussian PDF. The mean

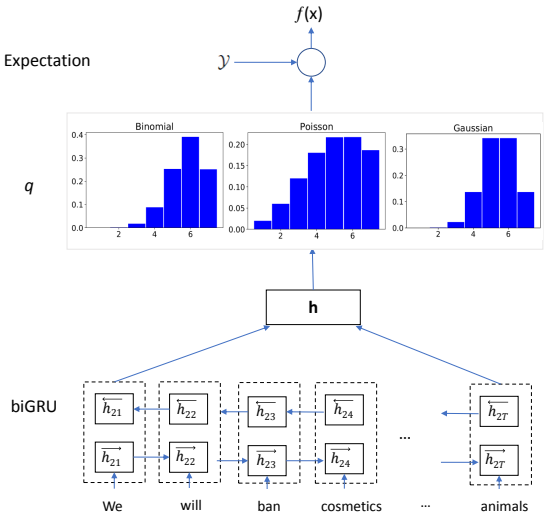


Figure 2: Illustration of the model architecture, comprising a biGRU over sentence tokens, to compute the parameter of one of the three distributions: p for Binomial (with $n = K-1$) and λ for Poisson. *Pmf* of these distributions are then used to define a categorical distribution over the ordinal classes $\mathcal{Y}$. 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,

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

This formulation allows efficient computation of divergence between p_* and q_x for optimization, which results in cross-entropy minimization ($\mathcal{L}_{\mathcal{D}\text{GAUSS}}$: 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
	$\text{biGRU}_{\text{REG}}$	1.83	0.47
	$\text{biGRU}_{\text{CLASS}}$	2.17	0.40
	$\text{biGRU}_{\text{REG}_{l_1}}$	1.99	0.46
Ordinal	$\text{biGRU}_{\text{CATEGORICAL}}$	1.80	0.48
Ordinal	$\text{biGRU}_{\text{BINOMIAL}}$	1.78	0.48
Ordinal	$\text{biGRU}_{\text{POISSON}}$	1.90	0.41
Ordinal	$\text{biGRU}_{\text{GAUSS}}$	1.72	**0.49**
Ordinal	$\text{biGRU}_{\text{GAUSS + CONTEXT}}$	**1.71**	**0.49**

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

mover's distance, $\text{EMD}(q_\theta(\mathcal{Y}|s),\ q_\omega(\mathcal{Y}|s))$; denoted as "$\mathcal{L}_{\text{UEMD}}$". EMD is defined as $\text{EMD}(q_\theta, q_\omega) = \frac{1}{K}^{\frac{1}{l}} \|\text{cmf}(q_\theta) - \text{cmf}(q_\omega)\|_l$, where $\text{cmf}(\cdot)$ 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).

$\text{biGRU}_{\text{REG}}$: biGRU model trained with a mean squared loss objective.

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

$\text{biGRU}_{\text{REG}_{l_1}}$: 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: $\text{biGRU}_{\text{GAUSS}}$ (Section 4.2.3), $\text{biGRU}_{\text{BINOMIAL}}$ (Section 4.2.1), or $\text{biGRU}_{\text{POISSON}}$ (Section 4.2.2). We also evaluate $\text{biGRU}_{\text{CATEGORICAL}}$, 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 ($\text{biGRU}_{\text{REG}}$) performs better than l_1 regression ($\text{biGRU}_{\text{REG}_{l_1}}$) and classification ($\text{biGRU}_{\text{CLASS}}$).

With respect to deep ordinal approaches, $\text{biGRU}_{\text{POISSON}}$ performs better than classification ($\text{biGRU}_{\text{CLASS}}$), but does not improve upon regression ($\text{biGRU}_{\text{REG}}$). The Binomial performs better than the Poisson, consistent with previous work (Beckham and Pal, 2017; da Costa et al., 2008). $\text{biGRU}_{\text{CATEGORICAL}}$ performs better than $\text{biGRU}_{\text{REG}}$, but not over unimodal approaches ($\text{biGRU}_{\text{BINOMIAL}}$, $\text{biGRU}_{\text{GAUSS}}$). 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 $\text{biGRU}_{\text{REG}}$. Adding context to $\text{biGRU}_{\text{GAUSS}}$ ($\text{biGRU}_{\text{GAUSS+CONTEXT}}$) provides a slight reduction in error.

5.1 Semi-supervised Learning

We next compare the performance of $\text{biGRU}_{\text{GAUSS+CONTEXT}}$ (SUP) and the semi-

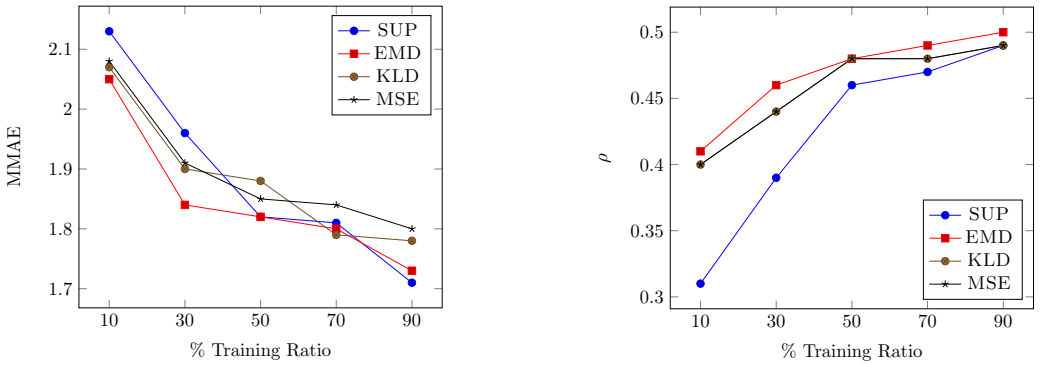


Figure 3: Prediction performance across different training ratios. Note that 90% = all the training data, as 10% is used for validation. The supervised ordinal model (SUP) and semi-supervised teacher–student models (MSE, KLD, EMD, given in Section 4.4) are compared on MMAE and Spearman's ρ.

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 $\lambda : \texttt{P(a)} \wedge \texttt{Q(a,b)} \rightarrow \texttt{R(b)}$, where P, Q, and R are predicates, a and 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
$\texttt{Manifesto(x)} \wedge \texttt{Policy(I)} \wedge \texttt{Specw(x, I)} \wedge \texttt{IdeologyMap(I, social left/right)} \rightarrow \texttt{socpos(x)}$ $\texttt{Manifesto(x)} \wedge \texttt{Policy(I)} \wedge \texttt{Specw(x, I)} \wedge \texttt{IdeologyMap(I, economic left/right)} \rightarrow \texttt{econpos(x)}$
Overall position — Model II
$\texttt{Manifesto(x)} \wedge \texttt{socpos(x)} \rightarrow \texttt{pos(x)}$ $\texttt{Manifesto(x)} \wedge \texttt{econpos(x)} \rightarrow \texttt{pos(x)}$
Global signals — Model III
$\texttt{Manifesto(x)} \wedge \texttt{Party(x, a)} \wedge \texttt{Manifesto(y)} \wedge \texttt{Party(y, b)} \wedge \texttt{Coalition(a, b)} \wedge \texttt{pos(x)} \rightarrow \texttt{pos(y)}$ $\texttt{Manifesto(x)} \wedge \texttt{Party(x, a)} \wedge \texttt{PreviousManifesto(x, t)} \wedge \texttt{Party(t, a)} \wedge \texttt{pos(t)} \rightarrow \texttt{pos(x)}$
Relative specificity — Model IV
$\texttt{Manifesto(x)} \wedge \texttt{Policy(I)} \wedge \texttt{SpecScale(x, I)} \wedge \texttt{IdeologyMap(I, social left/right)} \rightarrow \texttt{socpos(x)}$ $\texttt{Manifesto(x)} \wedge \texttt{Policy(I)} \wedge \texttt{SpecScale(x, I)} \wedge \texttt{IdeologyMap(I, economic left/right)} \rightarrow \texttt{econpos(x)}$

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 $|\text{Manifestos}| \times 57$ 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

We thank Robert Thomson for his inputs on the task definition. This work was funded in part by the Australian Government Research Training Program Scholarship, and the Australian Research Council.

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.

3.4 Election promises progress tracker [CL 2020]

Given the promises extracted from an election campaign (Section 3.3), the logical next step is to track a government's subsequent progress in terms of policy implementation. Towards this, political scientists have performed various promise progress tracking studies across countries, in a manual fashion. There are also news organizations whose journalists track progress manually, e.g., The Australian Broadcasting Corporation (ABC) promise tracker.[1] This requires a lot of manual effort, where here our goal is to automate this task, to help reduce the efforts of domain experts.

There are two major challenges to automate this analysis. First, hundreds of promises are made during an election, and not all the promises are tracked by journalists. Hence we have to first rank 'track-worthy' promises, which are based on several factors — specificity of promises, promise issue salience, and journalists' biases. Second, there are many possible evidence documents that are potentially relevant in tracking the status of a promise, and hence we must retrieve relevant evidence documents and classify the progress made. This is a non-trivial task. We curate a corpus by combining existing data annotated by journalists in news organizations across countries. Here, we face an immediate challenge to address — the progress status labels are not uniform across all the news organization trackers. For example, PolitiFact[2] (in the US political setting) has a *Compromise* status category which is not present in ABC's tracker. Hence we first align labels across existing tracking systems, and propose a two-stage automated solution — ranking track-worthy promises, and categorizing the promise status given evidence documents.

We use supervised learning approaches in both cases — track-worthiness estimation is a binary classification task given the promise text, and the progress status classification is posed as a three-way classification (irrelevant, positive or negative) given both the promise text and an evidence document. Building on the success of pre-trained language models we observed in our earlier work (Sections 3.2 and 3.3), we evaluate several other representation strategies including the bidirectional encoder representations from transformers (BERT) [Devlin et al. (2019)]. We observe that BERT performs best overall, and that the promise progress classification is harder than track-worthiness estimation. In the progress classification task, specifically, finding negative status is the hardest. This work is under submission to Computational Linguistics.

[1] https://www.abc.net.au/news/factcheck/promisetracker/
[2] https://www.politifact.com/truth-o-meter/promises/trumpometer

Towards Automatic Tracking of Election Promises

Shivashankar Subramanian*
The University of Melbourne

Trevor Cohn**
The University of Melbourne

Timothy Baldwin†
The University of Melbourne

An election promise is a commitment to future action made by a political party to the public, and delivering on the promises made is a core part of representative democracy. Political scientists have long studied how election promises translate into government programs and actual policy, as part of a broader study of governmental accountability, largely based on manual analysis. Many promises are made in the course of an election campaign, and keeping track of all the promises manually is a resource-intensive endeavor. This work is an attempt to develop a novel dataset for election promise tracking, and evaluate suitable language processing approaches to demonstrate that the task can be automated. We develop a novel dataset covering promises made by political parties in Australia, Canada, UK, and the US, and propose a two-stage approach: (1) selecting track-worthy promises from election campaigns; and (2) classifying progress made towards those promises, conditioned on evidence.

1. Introduction

Political science is a highly specialised and rich domain which encompasses a wide range of natural language processing challenges, rooted in the nature of political communication which contains multiple modalities of expression. This inter-disciplinary work melding political science and NLP is gaining popularity among NLP researchers, under the banner of computational social science, owing to the increasing availability of many digitized corpora. Some popular efforts toward curation and annotation of political text include the Comparative Manifesto Project (CMP) (Volkens et al. 2017) which contains a large collection of digitized manifestos, annotated for party policy positions. Policy Agendas Project is another major effort in political science (Kingdon 1984), which deals with analysis of policy agenda arising in the decision-making process. It includes text from a wide range of political actors (e.g., manifestos, bills, laws passed, news data, etc.), unlike CMP which is restricted to manifestos only. Other than these massive annotation efforts, the American Presidency project (Woolley and Peters 2011), is an example for curation-based work. It includes all the presidential documents — election debate transcripts, White House press briefings, proclamations, executive actions, etc. Presidential debates have been intensively studied from the Nixon-Kennedy contest of 1960 to the present (Isotalus 2011). Our work is a curation-based effort targeting the study of progress made towards election promises.

* Computing and Information Systems, The University of Melbourne, Australia. Email: shivashankar@student.unimelb.edu.au
** Computing and Information Systems, The University of Melbourne, Australia. Email: trevor.cohn@unimelb.edu.au
† Computing and Information Systems, The University of Melbourne, Australia. Email: tbaldwin@unimelb.edu.au

Update	Status	Time
President Donald Trump signed an executive order to withhold grant money from cities that protect undocumented immigrants, otherwise known as "sanctuary cities".	In Progress	Jan 2017
Attorney General Jeff Sessions said states and local jurisdictions pursuing Justice Department grants will need to prove compliance with federal laws in order to receive funds.	In Progress	Mar 2017
California judge halts Trump's executive order.	Stalled	Apr 2017
House passes bill to withhold certain federal grants from 'sanctuary cities'.	In Progress	Jun 2017
Federal appeals court says Donald Trump's order against sanctuary cities is unconstitutional.	Broken	Aug 2018

Table 1
Progress updates for the promise *Cancel all funding of sanctuary cities*. *In Progress* instances are updates with **Positive** progress, *Stalled* and *Broken* have **Negative** progress. Each update is based on one or more news articles, which we consider as evidence.

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 rhetorical functions fulfilled by an election campaign (Eder, Jenny, and Müller 2017; Subramanian, Cohn, and Baldwin 2019b), perhaps the most important is the contract they represent between parties and voters in terms of promises and prioritisation of political issues (Royed, Naurin, and Thomson 2019). Fulfilment of election pledges is central to the concept of representative democracy, and thus a keen area of focus by political scientists, in linking election promises to government programs and policies (Royed 1996; Thomson 2001; Schermann and Ennser-Jedenastik 2014). However research to date has been dominated by manual text analysis. With multiple news and related web sources, coupled with the need to keep track of promises over time, this takes significant expert time and effort. This motivates the need for automated election promise tracking systems. This task is analogous to fact-checking systems (Vlachos and Riedel 2014), which judge the veracity of a claim against a text corpus. Here, our goal is to determine whether the party forming government fulfills its promises or not. A major difference comes from the fact that progress made towards promises evolves over time. For example, the promise made by Donald Trump (Republican Party of US) to *Cancel all funding of sanctuary cities*[1] had many status updates as shown in Table 1. In this work, we propose a two-stage approach for tracking progress made towards delivering on promises, and create a new dataset for evaluating automatic promise tracking systems.

1.1 Two-stage approach

We propose a two-stage approach to promise tracking, with the following components:

1. **Track-worthiness estimation**: Since many promises are made during an election cycle, also with different levels of details and specificity (Subramanian, Cohn, and Baldwin 2019a), it is necessary to first select track-worthy promises. For instance, the ABC tracker

1 https://www.politifact.com/truth-o-meter/promises/trumpometer/promise/1400/cancel-all-funding-sanctuary-cities/

identified 78 promises made by the Liberal Party of Australia,[2] and PolitiFact (Trumpometer) covered 102 promises made by the US Republican Party,[3] both out of a much larger set of statements made during their respective election campaigns. In addition to the promise specific information, there can be other subjective components influencing the selection of promises tracked, and is beyond the scope of this focused study.

2. **Progress classification**: The next step is to track progress made towards those promises, using news and related sources as potential evidence. This includes both identifying suitable evidence, and classifying the status based on the evidence. We model it as a 3-way classification task, wherein a given piece of evidence is: (1) IRRELEVANT to a promise's progress; (2) indicative of POSITIVE progress towards the promise (partial or complete fulfilment); or (3) indicative of NEGATIVE progress towards the promise (stalled or broken).

Progress state classification must accommodate the dynamic nature of the progress towards fulfilling a promise, which is reflected in the status labels used across promise tracking websites (Full-Fact 2020). For example, the ABC has 'Kept', 'Broken', 'In Progress', and 'Stalled', whereas Polimetre has only 'Kept', 'Broken' and 'In Progress'. In this work we simplify and consolidate these labels inventories into a 3-way classification scheme, capturing whether the evidence reports on progress being made, and when it does is it positive or negative.

The major contributions of this work are as follows:

- To the best of our knowledge this is the first attempt towards formulating the task of an automated election promise progress tracker.
- We develop and release a dataset related to promises from political parties of Australia, Canada, UK, and the US (available at https://www.github.com/shivashankarrs/promise-tracker).
- We deploy and evaluate various approaches over the task, as strong baselines for future work.

2. Related Work

Political scientists have advocated that strong program-to-policy linkage is central to the mandate theory of democracy and the responsible party model (Downs 1957; Thomson 2001; Royed, Naurin, and Thomson 2019). Various factors such as single-party versus coalition-based government (e.g., US vs. Germany), ideological position, issue salience, and institutional constraints have been shown to influence pledge fulfilment (Thomson et al. 2017; Pétry and Duval 2018; Klingemann, Hofferbert, and Budge 1994). Several organizations in different countries have developed promise trackers, where journalists verify the progress of promises using reliable sources as evidence (Full-Fact 2020), all based on manual checking (see also Section 3). Given the scale of web sources and the dynamic nature of political reporting and a government's progress towards delivering promises, there are obvious benefits to automating the task.

The closest automated task is fact-checking (Vlachos and Riedel 2014; Thorne and Vlachos 2018), which deals with verifying the truthfulness of claims. A popular application is checking the truthfulness of claims made during political campaigns (Hassan et al. 2017). Here, the first step is to find check-worthy claims from campaign text, e.g., in US presidential and vice-presidential debates (Hassan, Li, and Tremayne 2015; Vasileva et al. 2019). There was also a shared task targeting the challenge of identifying check-worthy political claims and verifying

2 https://www.abc.net.au/news/factcheck/promisetracker/
3 https://www.politifact.com/truth-o-meter/promises/trumpometer/

Organization	Term start	Party	#Promises	Country
ABC	2013	Liberal Party	78	Australia
Polimetre	2015	Liberal Party	353	Canada
GovTracker	2015 and 2017	Conservatives	431	UK
PolitiFact	2008 and 2012	Democrats	533	US
PolitiFact	2016	Republicans	102	US

Table 2
Statistics of promises tracked by different organizations. "Term start" denotes the election year during which the Party made the promises. Note that only the promises made by the party which forms government are tracked.

Class	Promise	Party	Election
TRACKED	We will increase the Inheritance Tax zero tax threshold per couple to £1 million by 2020.	The Conservative Party, UK	2015
TRACKED	Will amend strike action legislation so that strikes must have more support before they are considered legal.	The Conservative Party, UK	2015
NOT TRACKED	We will build more modern infrastructure to get things moving with an emphasis on reducing the bottlenecks on gridlocked roads and highways.	The Liberal Party of Australia	2013
NOT TRACKED	We will provide targeted support for young people between the ages of 18 and 24 so that everyone is given the very best chance of getting into work.	The Conservative Party, UK	2017

Table 3
Example TRACKED and NOT TRACKED promises, party which makes the promise and the election details are given.

their truthfulness at CLEF-2018 (Nakov et al. 2018). In this work we both find track-worthy promises, and classifying the status of those promises using suitable evidence. Key differences over fact-checking of political speeches are: the status of a promise evolves over time; and also a promise tracker deals with tracking post-election accountability for promises made during an election campaign. Perhaps the most closely-related research is our earlier work on identifying promises and categorizing their specificity from election campaign text (Subramanian, Cohn, and Baldwin 2019a). In this work, we deal with the subsequent steps of ranking promises for track-worthiness, and tracking their progress.

Other related work include analyzing truthfulness of news and social media posts (Castillo, Mendoza, and Poblete 2011; Popat et al. 2017), as well as studying credibility, bias, and propaganda in news media sources (Zhang et al. 2019; Baly et al. 2018).

3. Dataset

We constructed a dataset covering seven governments across four countries, as shown in Table 2.

Given Status	Class	Organization
In Progress/In the works	POSITIVE	ABC, POL, PF
Kept in parts	POSITIVE	POL
Stalled	NEGATIVE	ABC, POL
Broken	NEGATIVE	ABC, GOV, POL, PF
Kept	POSITIVE	ABC, POL, PF
Completed	POSITIVE	GOV
Completed (Maintaining)	POSITIVE	GOV
Compromise	POSITIVE/NEGATIVE	PF

Table 4
Mapping from status categories to our classification for different promise tracker organizations (ABC, GOV=GovTracker, POL=Polimetre, PF=PolitiFact). The *Compromise* category is re-annotated manually, as described in Section 3. GovTracker also has a *Pending* category, for promises that have no action from government (i.e., no evidence), which is ignored in our analysis.

3.1 Track-worthy Promises

For the first task of track-worthiness ranking, we consider the promises tracked by the news organizations listed in Table 2 as positive instances (1497 in total). Using the election manifesto text from those parties (Volkens et al. 2017), we manually select promises that are not tracked, and consider these as negative instances. We create a roughly balanced dataset of TRACKED and NOT TRACKED promises,[4] with a total of over 3000 promises (1497 positive and 1503 negative instances, see Table 3 for examples).

3.2 Progress Classification

For progress classification, we curate the promise status categorization and corresponding evidence as found in our primary source (where evidence is from news and related sources). We map the different statuses across promise tracker organizations into POSITIVE and NEGATIVE categories, as given in Table 4. For the COMPROMISE class, we manually categorize the evidence as either POSITIVE or NEGATIVE. Inter-annotator agreement for classifying COMPROMISE cases, was $\kappa = 0.71$ (substantial agreement), measured between two annotators over 10% of the samples. The remaining cases were annotated without redundancy. We removed those promises which had no update or evidence, leaving a total of 1201 promises. Since retrieving relevant evidence is a non-trivial task, we also include an additional IRRELEVANT class. For each promise, we randomly sample evidence from other promises to serve as IRRELEVANT cases, after manually verifying their irrelevance. We keep the ratio between the RELEVANT and IRRELEVANT classes balanced, since there is no known distribution for irrelevant cases, giving 5850 promise–evidence document pairs — 2347 (POSITIVE), 554 (NEGATIVE) and 2949 (IRRELEVANT). Note that, NEGATIVE instances are less than 20% among RELEVANT cases (including POSITIVE) and less than 10% overall. Content was scraped from the linked evidence documents online, and for those documents (news reports and press-releases largely) which were no longer accessible, the Internet Archive was used to recover the original.[5] The evidence collection contains 2650 documents from 189 sources, comprising over 143k sentences and 3.2M tokens.

4 We opted for a balanced distribution as the natural distribution of tracked and untracked promises is unclear, given that there is no explicit record on the part of promise tracking organizations as to what is *not* tracked.
5 https://archive.org/web

Approach	F1	ROC AUC
Random	0.49	0.50
NB	0.52	0.53
LDA	0.56	0.57
MLP	0.61	0.65
MLP-Glove	0.62	0.67
CNN-Glove	0.63	0.68
USE	0.66	0.73
ELMo	0.69	0.77
BERT	**0.73**	**0.84**

Table 5
Track-worthiness estimation results, showing the F_1 score and the area under the receiver operating characteristics curve using five-fold cross-validation; best results in bold.

4. Evaluation

For the first task of track-worthiness estimation, we evaluate various approaches using the promise text as input (described in Section 4.1). For the promise progress classification task, we take as input both the promise sentence and evidence text. The proposed task can be seen as an instance of natural language inference (NLI; Bowman et al. (2015)), in which the evidence sentences and promise text serve as the premise and hypothesis, respectively. We evaluate several approaches, including leading NLI methods, and report the results in Section 4.2.

4.1 Track-worthiness Estimation

Given the promise text, we use the following baseline approaches for track-worthiness estimation, to benchmark the complexity of the task. Results are presented in Table 5, based on F1-score and area under the Receiver Operating Characteristic curve (AUC), which measures how well positive instances are ranked above negative ones, with AUC = 0.5 indicating a random ordering of positive and negative instances.

Approach. First we evaluate a RANDOM prediction baseline, noting that it is a binary classification task over a balanced dataset. Then we evaluate traditional machine learning approaches with a term-frequency unigram sentence representation. The approaches include, Naive Bayes (NB), Linear Discriminant Analysis (LDA), and ADABOOST with decision tree base classifiers.

Since non-linear approaches have been shown to model text better than linear techniques, we use a multi-layer perceptron (MLP) model, with term-frequency unigram input representation, a single hidden layer with a ReLU activation, and a logistic output layer. We additionally evaluate a MLP model with a single hidden layer and a logistic output layer, but an input representation based on average pooling over 300-d GloVe embeddings (MLP-GLOVE) (Pennington, Socher, and Manning 2014).

Though MLPs are universal function approximators, they cannot capture sequential information in text. We thus also implement a convolutional neural network model, with unigram, bigram and trigram filters (20 each, with ReLU activations), using GloVe word representations, passed through an intermediate hidden layer, and a logistic output layer (CNN-GLOVE).

In recent years, language model-based text representations have gained popularity, motivated by which, we evaluate different contextual word representation strategies. First, we use Universal

Approach	Micro-F1	Macro-F1
Majority	0.34	0.22
Promise Only	0.32	0.23
LR	0.36	0.27
MLP-GloVe$_{NLI}$	0.34	0.25
CNN-GloVe$_{NLI}$	0.35	0.29
USE$_{NLI}$	0.50	0.39
ELMo$_{NLI}$	0.42	0.33
BERT$_{NLI}$	**0.59**	**0.46**

Table 6
Promise progress classification results based on F_1 score, measured over three classes (**POSITIVE,** **NEGATIVE** and **IRRELEVANT**); best results in bold.

Sentence Embeddings (USE) (Cer et al. 2018) to encode the input sentences,[6] passed through a hidden layer and a logistic output layer. Next, we use deep contextualized word representations (ELMO) (Peters et al. 2018) to encode the input sentences,[6] passed through a logistic output layer. Finally, we evaluate deep bidirectional transformers (BERT) (Devlin et al. 2019), based on a [CLS] sentence encoding in the final layer,[6] passed through a logistic output layer. All layers of BERT are fine-tuned with respect to the task objective.

4.2 Promise Progress Classification

Promises are conditioned on the evidence text for their progress status classification. Specifically, the top-5 most similar sentences are retrieved from the provided evidence document based on the promise sentence, using a term-frequency unigram representation and cosine similarity. This strategy has been shown to be effective for other similar tasks (Chakrabarty, Alhindi, and Muresan 2018). We evaluate the following models based on cross-validated micro- and macro-average F1-scores. The results are provided in Table 6.

Approaches. First we evaluate a MAJORITY class baseline, noting that the dataset is not balanced, where the distribution of POSITIVE and NEGATIVE classes are based on the promise tracker data from news organizations, and NEGATIVE cases are less than 10% overall. Next, we evaluate a PROMISE ONLY model, with a term-frequency unigram promise text representation (and no evidence representation), and logistic regression for prediction, as a means of measuring hypothesis bias in the data and quantifying dataset artefacts. We also evaluate a logistic regression (LR) model over concatenated term-frequency unigram representations of each of the promise and evidence text.

Similar to the track-worthiness estimation task, we evaluate a range of non-linear models, where promise and evidence text are represented using MLP, CNN, USE, ELMo and BERT, in conjunction with the standard NLI model (Bowman et al. 2015). We evaluate an MLP with average pooling over GloVe word embeddings for the promise and evidence, concatenated and passed through a hidden layer, followed by a three-way softmax output layer (MLP-GLOVE$_{NLI}$). Again within an NLI setup, we use a CNN model, with unigram, bigram and trigram filters

6 Details of the pre-trained encoding models: USE encodings are 512d, ELMo encodings are 1024d, for BERT we use uncased model with 12 layers and 768d embeddings.

(20 in total) over GloVe word embeddings to represent the promise and evidence text (CNN-GLOVE$_{NLI}$).

Additionally, we evaluate language model-based representations, starting with USE to encode both the promise and evidence text (average pooling over the five evidence sentences), concatenated and passed through a hidden layer, and finally a three-way softmax output layer (USE$_{NLI}$). We similarly implement a classifier using ELMo embeddings in place of USE (ELMO$_{NLI}$). Finally, we use BERT [CLS] encoding based on concatenating the promise and evidence text pair with a [SEP] delimiter, and otherwise uses the same neural architecture (BERT$_{NLI}$).

4.3 Analysis

From Tables 5 and 6, we can see that approaches using contextual text representations (USE, ELMo, and BERT) outperform methods using GloVe, as well as term-frequency representations. The BERT-based models perform best over both tasks.

For the track-worthiness estimation task, to understand where the BERT model fails, we analyse the relation to the specificity of the promise, based on the hypothesis that less specific promises are less likely to be tracked (Subramanian, Cohn, and Baldwin 2019a). From the validation set, we sampled 20% of instances where the model made mistakes (both false positives and false negatives). We analyzed the specificity of the promise text manually, and observed that over 76% of false positive instances were specific, and over 60% of false negatives were vague promises, which shows that the model relies on specificity as a cue. For the promise progress classification task, the F1 performance for IRRELEVANT is 0.73, POSITIVE is 0.53, and NEGATIVE is 0.13. Overall, promise progress (positive and negative) classification is harder than relevance classification (identifying irrelevant evidence) and also track-worthiness estimation. Predicting negative instances is also harder than the other cases, due to the class imbalance (NEGATIVE comprises $<$10% of the dataset).

5. Conclusion and Future Work

We have proposed a novel two-stage approach for tracking progress made towards election promises, constructed a dataset covering four countries, and provided experimental results using various language processing models. Though we found that specificity of promises influences its track-worthiness, there are other factors to investigate including issue salience and journalist bias, which deserve further attention. Overall, the results are encouraging and provide a baseline for future research in this area, but also point to open challenges such as improving the performance of predicting negative progress. Finally, both the track-worthiness estimation and promise progress detection tasks assume the promises to be independent instances, but the automated approaches can benefit from modeling the dependencies between promises.

Acknowledgement

We thank Jenny Lewis, Aaron Martin and Robert Thomson for discussions and inputs on this task. This work was funded in part by the Australian Government Research Training Program Scholarship, and the Australian Research Council.

References

Baly, Ramy, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, ACM.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chakrabarty, Tuhin, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 127–131.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Downs, Anthony. 1957. An economic theory of political action in a democracy. *Journal of political economy*, 65(2):135–150.

Eder, Nikolaus, 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.

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

Full-Fact. 2020. Full Fact document: The Future of Promise Tracking. https://bit.ly/37K1FkL. Accessed: 2020-01-04.

Hassan, Naeemul, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management*, pages 1835–1838, ACM.

Hassan, Naeemul, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Isotalus, Pekka. 2011. Analyzing presidential debates. *Nordicom Review*, 32(1):31–43.

Kingdon, John W. 1984. Agendas, alternatives and public policies. In *The Oxford handbook of classics in public policy and administration*, volume 45.

Klingemann, H.D., R.I. Hofferbert, and I. Budge. 1994. *Parties, policies, and democracy*. Theoretical lenses on public policy. Westview Press.

Nakov, Preslav, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 372–387, Springer.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, Matthew E, 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, NAACL-HLT 2018*, pages 2227–2237.

Pétry, François and Dominic Duval. 2018. Electoral promises and single party governments: the role of party ideology and budget balance in pledge fulfillment. *Canadian Journal of Political Science/Revue canadienne de science politique*, 51(4):907–927.

Popat, Kashyap, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Royed, Terry J. 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.

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

Schermann, Katrin and Laurenz Ennser-Jedenastik. 2014. Coalition policy-making under constraints: Examining

the role of preferences and institutions. *West European Politics*, 37(3):564–583.
Subramanian, Shivashankar, Trevor Cohn, and Timothy Baldwin. 2019a. Deep ordinal regression for pledge specificity prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 1729–1740.
Subramanian, Shivashankar, Trevor Cohn, and Timothy Baldwin. 2019b. 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.
Thomson, Robert. 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.
Thomson, Robert, 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.
Thorne, James and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.
Vasileva, Slavena, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. *arXiv preprint arXiv:1908.07912*.
Vlachos, Andreas and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.
Volkens, Andrea, 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.
Woolley, John T and Gerhard Peters. 2011. The American Presidency Project. Santa Barbara, CA.
Zhang, Yifan, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, et al. 2019. Tanbih: Get to know what you are reading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 223–228.

3.5 Online petitions popularity prediction [ACL 2018]

All our previous work (Sections 3.1 – 3.4) focused on text generated by political parties and their representatives. Here we look at the primary stakeholder, the *voters*, for and by whom a democracy functions. Political parties present their policy proposals during the election campaign, and bring them to functioning via several mechanisms including legislation, after the election is over. Voters already play a major role by voting, which can seen as some form of feedback based on both the policy proposal, and policy implementation of political parties and their representatives in the previous government. Though voting allows voter involvement in the process, this can be done only once every election cycle, and does not allow active voter engagement in the policy-making process. To mitigate this, one important instrument is *online petitions*, especially those implemented by the government of a country, which allow people to connect with their representatives. On such platforms, petition authors are responsible for both writing the petition as well as promoting them to gather support from other citizens. At the same time, the platform may have a moderation team to work with the petition authors and support them to write persuasive petitions. Secondly, the platform coordinators make decisions that can impact the reach of petitions, for example, which petition to promote on their homepage or which petition to delete after a certain period of time.

In this work, we want to predict the popularity of a petition as given by the number of voters who support the change or request. We propose to model the popularity prediction using petition's text content, which can directly evaluate their persuasiveness. More importantly, since fate of petitions is decided during the early few days [Yasseri et al. (2017)], the model can be used to predict popularity at the time of submission which can help both the petition owners as well as the platform moderators to improve their campaign strategies, and thereby to enhance public engagement. This can also be seen as a way to promote *voter advocacy*. The main motivation for this task is that active participation of the public, in tandem with political parties fulfilling their duties, is essential for a healthy democracy.

We model the regression task using a convolutional neural network, and we augment the model with a set of hand-crafted features. We observe that there is no improvement in performance with the additional hand-crafted features. Following that, in order to understand what patterns are captured by the deep learning model, we compute the dependency between hidden representations (independent variables) and hand-crafted features (dependent variables). We find that the deep learning model captures many syntactic and semantic features, but does not seem to

capture all the hand-crafted features, especially the domain related ones such as political bias of the petition text or policy-area popularity computed based on the previous election manifestos. This work was published at ACL 2018.

Content-based Popularity Prediction of Online Petitions Using a Deep Regression Model

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

Abstract

Online petitions are a cost-effective way for citizens to collectively engage with policy-makers in a democracy. Predicting the popularity of a petition — commonly measured by its signature count — based on its textual content has utility for policy-makers as well as those posting the petition. In this work, we model this task using CNN regression with an auxiliary ordinal regression objective. We demonstrate the effectiveness of our proposed approach using UK and US government petition datasets.[1]

1 Introduction

A petition is a formal request for change or an action to any authority, co-signed by a group of supporters. Research has shown the impact of online petitions on the political system (Lindner and Riehm, 2011; Hansard, 2016; Bochel and Bochel, 2017). Modeling the factors that influence petition popularity — measured by the number of signatures a petition gets — can provide valuable insights to policy makers as well as those authoring petitions (Proskurnia et al., 2017).

Previous work on modeling petition popularity has focused on predicting popularity growth over time based on an initial popularity trajectory (Hale et al., 2013; Yasseri et al., 2017; Proskurnia et al., 2017), e.g. given the number of signatures a petition gets in the first x hours, prediction of the total number of signatures at the end of its lifetime. Asher et al. (2017) and Proskurnia et al. (2017) examine the effect of sharing petitions on Twitter on its overall success, as a time series regression task. Other work has analyzed the importance of content on the success of the petition (Elnoshokaty et al., 2016). Proskurnia et al. (2017) also consider the anonymity of authors and petitions featured on the front-page of the website as additional factors. Huang et al. (2015) analyze 'power' users on petition platforms, and show their influence on other petition signers.

In general, the target authority for a petition can be political or non-political. In this work, we use petitions from the official UK and US government websites, whereby citizens can directly appeal to the government for action on an issue. In the case of UK petitions, they are guaranteed an official response at 10k signatures, and the guarantee of parliamentary debate on the topic at 100k signatures; in the case of US petitions, they are guaranteed a response from the government at 100k signatures. Political scientists refer to this as *advocacy democracy* (Dalton et al., 2003), in that people are able to engage with elected representatives directly. Our objective is to predict the popularity of a petition at the end of its lifetime, solely based on the petition text.

Elnoshokaty et al. (2016) is the closest work to this paper, whereby they target Change.org petitions and perform correlation analysis of popularity with the petition's category, target goal set,[2] and the distribution of words in General Inquirer categories (Stone et al., 1962). In our case, we are interested in the task of automatically predicting the number of signatures.

We build on the convolutional neural network (CNN) text regression model of Bitvai and Cohn (2015) to infer deep latent features. In addition, we evaluate the effect of an auxiliary ordinal regression objective, which can discriminate petitions that attract different scales of popular-

[1] The code and data from this paper are available from http://github.com/shivashankarrs/Petitions

[2] See http://bit.ly/2BXd0Sl.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 182–188
Melbourne, Australia, July 15 - 20, 2018.

ity (e.g., 10 signatures, the minimum count needed to not be closed vs. 10k signatures, the minimum count to receive a response from UK government).

Finally, motivated by text-based message propagation analysis work (Tan et al., 2014; Piotrkowicz et al., 2017), we hand-engineer features which capture wording effects on petition popularity, and measure the ability of the deep model to automatically infer those features.

2 Proposed Approach

Inspired by the successes of CNN for text categorization (Kim, 2014) and text regression (Bitvai and Cohn, 2015), we propose a CNN-based model for predicting the signature count. An outline of the model is provided in Figure 1. A petition has three parts: (1) title, (2) main content, and (3) (optionally) additional details.[3] We concatenate all three parts to form a single document for each petition. We have n petitions as input training examples of the form $\{a_i, y_i\}$, where a_i and y_i denote the text and signature count of petition i, respectively. Note that we log-transform the signature count, consistent with previous work (Elnoshokaty et al., 2016; Proskurnia et al., 2017).

We represent each token in the document via its pretrained GloVe embedding (Pennington et al., 2014), which we update during learning. We then apply multiple convolution filters with width one, two and three to the dense input document matrix, and apply a ReLU to each. They are then passed through a max-pooling layer with a $\tanh$ activation function, and finally a multi-layer perceptron via the exponential linear unit activation,

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha\left(\exp(x) - 1\right) & \text{otherwise}, \end{cases}$$

to obtain the final output (y_i), which is guaranteed to be positive. We train the model by minimizing mean squared error in log-space,

$$\mathcal{L}_{reg} = \frac{1}{n}\sum_{i=1}^{n} \|\hat{y}_i - y_i\|_2^2 , \qquad (1)$$

where $\hat{y}_i$ is the estimated signature count for petition i. We refer to this model as $\mathsf{CNN}_{\mathsf{regress}}$.

2.1 Auxiliary Ordinal Regression Task

We augment the regression objective with an ordinal regression task, which discriminates petitions

[3] Applicable for the UK government petitions only.

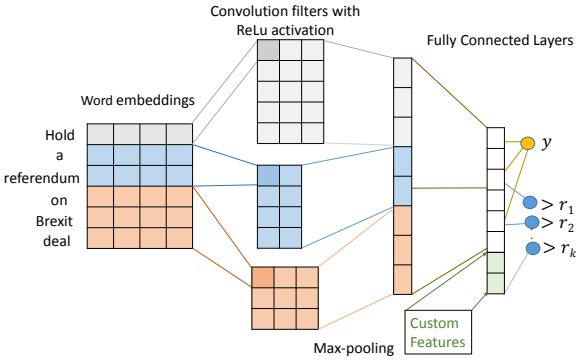


Figure 1: CNN-Regression Model. y denotes signature count. $> r_k$ is the auxiliary task that denotes p(petition attracting $> r_k$ signatures).

that achieve different scale of signatures. The intuition behind this is that there are pre-determined thresholds on signatures which trigger different events, with the most important of these being 10k (to guarantee a government response) and 100k (to trigger a parliamentary debate) for the UK petitions; and 100k (to get a government response) for the US petitions. In addition to predicting the number of signatures, we would like to be able to predict whether a petition is likely to meet these thresholds, and to this end we use the exponential ordinal scale based on the thresholds $O = \{10, 100, 1000, 10000, 100000\}$.[4] Overall this follows the exponential distribution of signature counts closely (Yasseri et al., 2017).

We transform the ordinal regression problem into a series of simpler binary classification subproblems, as proposed by Li and Lin (2007). We construct binary classification objectives for each threshold in O. For each petition i we construct an additional binary vector $\vec{o}_i$, with a 0–1 encoding for each of the ordinal classes ($\{a_i, y_i, \vec{o}_i\}$). Note that the transformation is done in a consistent way, i.e., if a petition has y signatures, then in addition to immediate lower-bound threshold in O determined by $l = \lfloor \log_{10} y \rfloor$ (for $y < 10^6$), all classes which have a lesser threshold are also set to 1 ($o_{t:t<l}$).

With this transformation, apart from the real-valued output y_i, we also learn a mapping from $\mathbf{h}_i$ with sigmoid activation for each class ($\vec{r}_i$). Finally we minimize cross-entropy loss for each binary classification task, denoted $\mathcal{L}_{aux}$.

[4] We use $O = \{1000, 10000, 100000\}$ for the US petitions, as only petitions which get a minimum of 150 signatures are published on the website.

Overall, the loss function for the joint model is:

$$\mathcal{L}_J = \mathcal{L}_{reg} + \gamma \mathcal{L}_{aux} \quad (2)$$

where $\gamma \geq 0$ is a hyper-parameter which is tuned on the validation set. We refer to this model as $\text{CNN}_{\text{regress+ord}}$.

3 Hand-engineered Features

We hand-engineered custom features, partly based on previous work on non-petition text. This includes features from Tan et al. (2014) and Piotrkowicz et al. (2017) such as structure, syntax, bias, polarity, informativeness of title, and novelty (or freshness), in addition to novel features developed specifically for our task, such as policy category and political bias features. We provide a brief description of the features below:

- Additional Information (ADD): binary flag indicating whether the petition has additional details or not.
- Ratio of indefinite (IND) and definite (DEF) articles.
- Ratio of first-person singular pronouns (FSP), first-person plural pronouns (FPP), second-person pronouns (SPP), third-person singular pronouns (TSP), and third-person plural pronouns (TPP).
- Ratio of subjective words (SUBJ) and difference between count of positive and negative words (POL), based on General Inquirer lexicon.
- Ratio of biased words (BIAS) from the bias lexicon (Recasens et al., 2013).
- Syntactic features: number of nouns (NNC), verbs (VBC), adjectives (ADC) and adverbs (RBC).
- Number of named entities (NEC), based on the NLTK NER model (Bird et al., 2009).
- Freshness (FRE): cosine similarity with all previous petitions, inverse weighted by the difference in start date of petitions (in weeks).
- Action score of title (ACT): probability of title conveying the action requested. Predictions are obtained using an one-class SVM model built on the universal representation (Conneau et al., 2017) of titles of rejected petitions,[5] as they don't contain any action request. These rejected petitions are not part of our evaluation dataset.
- Policy category popularity score (CSC): commonality of the petition's policy issue (Subramanian et al., 2017), based on the recent UK/US election manifesto promises.
- Political bias and polarity: relative leaning/polarity based on: (a) $\frac{\#\text{left}+\#\text{right}}{\#\text{left}+\#\text{right}+\#\text{neutral}}$ (PBIAS) (b) $\frac{\#\text{left}-\#\text{right}}{\#\text{left}+\#\text{right}}$ (L–R). Sentence-level *left*, *right* and *neutral* classes are obtained using a model built on the CMP dataset, and the categorization given by Volkens et al. (2013).

The custom features are passed through a hidden layer with $\tanh$ activations ($\mathbf{c}_i$), and concatenated with the hidden representation learnt using the dense input document (Section 2), $\begin{bmatrix} \mathbf{h}_i \\ \mathbf{c}_i \end{bmatrix}$, before mapping to the output layer (Figure 1). We refer to this model as $\text{CNN}_{\text{regress+ord+feat}}$. We use the Adam optimizer (Kingma and Ba, 2014) to train all our models.

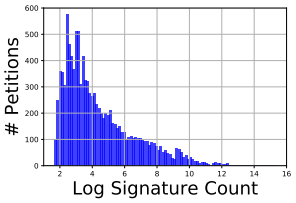


Figure 2: UK Petitions Signature Distribution.

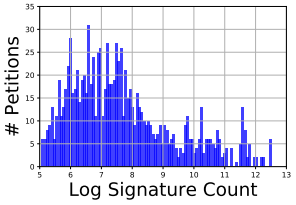


Figure 3: US Petitions Signature Distribution

4 Evaluation

We collected our data from the UK[6] and US[7] government websites over the term of the 2015–17 Conservative and 2011–14 Democratic governments respectively. The UK dataset contains 10950 published petitions, with over 31m signatures in total. We removed US petitions with $\leq$ 150 signatures, resulting in a total of 1023 petitions, with over 12m signatures in total. We split the data chronologically into train/dev/test splits based on a 80/10/10 breakdown. Distribution over log signature counts is given in Figures 2 and 3.

To analyze the statistical significance of each feature varying across ordinal groups O, we ran

[5] https://petition.parliament.uk/help

[6] https://petition.parliament.uk

[7] https://petitions.whitehouse.gov/

a Kruskal-Wallis test (at $\alpha = 0.05$: Kruskal and Wallis (1952)) on the training set. The test results in the test statistic H and the corresponding p-value, with a high H indicating that there is a difference between the two groups. The analysis is given in Table 2, where $p < 0.001$, $p < 0.01$ and $p < 0.05$ are denoted as "***", "**" and "*", respectively. Note that the ordinal groups are different for the two datasets: analyzing the UK dataset with the same ordinal groups used for the US dataset ({1000,10000,100000}) resulted in a similarly sparse set of significance values for non-syntactic features as the US dataset.

We benchmark our proposed approach against the following baseline approaches:

Mean: average signature count in the raining set.

Linear$_{\text{BoW}}$: linear regression (Linear) model using TF-IDF weighted bag-of-words features.

Linear$_{\text{GI}}$: linear regression model based on word distributions from the General Inquirer lexicon; similar to (Elnoshokaty et al., 2016), but without the target goal set or category of the petition (neither of which is relevant to our datasets).

SVR$_{\text{BoW}}$: support vector regression (SVR) model with RBF kernel and TF-IDF weighted bag-of-words features.

SVR$_{\text{feat}}$: SVR model using the hand-engineered features from Section 3.

SVR$_{\text{BoW+feat}}$: SVR model using combined TF-IDF weighted bag-of-words and hand-engineered features.

We present the regression results for the baseline and proposed approaches based on: (1) mean absolute error (MAE), and (2) mean absolute percentage error (MAPE, similar to Proskurnia et al. (2017)), calculated as $\frac{100}{n}\sum_{i=1}^{n}\frac{|\hat{y_i}-y_i|}{y_i}$. Results are given in Table 1.

The proposed CNN models outperform all of the baselines. Comparing the CNN model with regression loss only, CNN$_{\text{regress}}$, and the joint model, CNN$_{\text{regress+ord}}$ is superior across both datasets and measures. When we add the hand-engineered features (CNN$_{\text{regress+ord+feat}}$), there is a very small improvement. In order to further understand the effect of the hand-engineering features without the ordinal regression loss, we use it only with the regression task (CNN$_{\text{regress+feat}}$), which mildly improves over CNN$_{\text{regress}}$, but is below CNN$_{\text{regress+ord+feat}}$. We also evaluate a variant of CNN$_{\text{regress+ord+feat}}$ with an additional hidden layer, given in the final row of Table 1, and find it to lead to further improvements in the regression results. Adding more hidden layers did not show further improvements.

4.1 Classification Performance

The F-score is calculated over the three classes of $[0, 10000)$, $[10000, 100000)$ and $[100000, \infty)$ (corresponding to the thresholds at which the petition leads to a government response or parliamentary debate) for the UK dataset; and $[150, 100000)$ and $[100000, \infty)$ for the US dataset, by determining if the predicted and actual signature counts are in the same bin or not. We also built an SVM-based ordinal classifier (Li and Lin, 2007) over the significant ordinal classes, as an additional baseline. The CNN models struggle to improve F-score (in large part due to the imbalanced data). For the UK dataset, CNN models with an ordinal objective (CNN$_{\text{regress+ord}}$ and CNN$_{\text{regress+ord+feat}}$) result in a macro-averaged F-score of 0.36, compared to 0.33 for all other methods. But for the US dataset, which is a binary classification task, all methods obtain a 0.49 F-score. In addition to text, considering other factors such as early signature growth (Hale et al., 2013) — which determines the timeliness to get the issue online on the US website — could be necessary.

4.2 Latent vs. Hand-engineered Features

Finally, we built a linear regression model with the estimated hidden features from CNN$_{\text{regress+ord}}$ as independent variables and hand-engineered features as dependent variables, to study their linear dependencies in a pair-wise fashion. The most significant dependencies (given by p-value, p_{hidden}) over the test set are given in Table 2. We found that the model is able to learn latent feature representations for syntactic features (NNC, VBC, ADC,[8] RBC[9]), FRE, NEC, IND and DEF,[8] but not the other features — these can be considered to provide deeper information than can be extracted automatically from the data, or else information that has no utility for the signature prediction task. From the analysis in Table 2, some of the features that vary across ordinal groups are not linearly dependent with the deep latent features. These include ADD,[8] BIAS, CSC,[8] PBIAS,[9] and L–R, where the latter ones are policy-related features. This indicates that the custom features and hidden features contain complementary signals.

[8] UK dataset only.

[9] US dataset only.

Approach	UK Petitions MAE	UK Petitions MAPE	US Petitions MAE	US Petitions MAPE
Mean	4.37	159.7	2.82	44.61
$\text{Linear}_{\text{BoW}}$	1.75	57.56	2.51	37.01
$\text{Linear}_{\text{GI}}$	1.77	58.22	1.84	27.71
SVR_{BoW}	1.53	45.35	1.39	20.37
SVR_{feat}	1.54	46.96	1.40	20.48
$\text{SVR}_{\text{BoW+feat}}$	1.52	44.71	1.39	20.38
$\text{CNN}_{\text{regress}}$	1.44	36.72	1.24	14.98
$\text{CNN}_{\text{regress+ord}}$	1.42	33.86	1.22	14.68
$\text{CNN}_{\text{regress+ord+feat}}$	1.41	32.92	1.20	14.47
$\text{CNN}_{\text{regress+feat}}$	1.43	35.84	1.23	14.75
$\text{CNN}_{\text{regress+ord+feat}}$ + Additional hidden layer	**1.40**	**31.68**	**1.16**	**14.38**

Table 1: Results over UK and US Government petition datasets. Best scores are given in bold.

Feature	Description	UK Petitions H	UK Petitions p	UK Petitions p_{hidden}	US Petitions H	US Petitions p	US Petitions p_{hidden}
ADD	Additional details	94.59	***				
IND	Indefinite articles	14.87	*		8.56	*	*
DEF	Definite articles	34.91	***	*	3.69		
FSP	First-person singular pronouns	53.36	***		6.84	*	
FPP	First-person plural pronouns	11.26	*		6.10		
SPP	Second-person pronouns	13.80	*		3.95		
TSP	Third-person singular pronouns	5.82			9.07	*	
TPP	Third-person plural pronouns	16.13	**		5.58		
SUBJ	Subjective words	12.25	*		7.21	*	***
POL	Polarity	2.60		*	4.27		
BIAS	Biased words	11.92	*		4.56	*	
NNC	Nouns	7.34		***	1.93		**
VBC	Verbs	2.75		**	7.46	*	***
ADC	Adjectives	26.14	***	***	4.07		
RBC	Adverbs	17.09	**		2.99		*
NEC	Named entities	51.11	***	***	3.94		*
FRE	Freshness	86.97	***	*	13.86	**	*
ACT	Title's action score	3.89			3.54		
CSC	Policy category popularity	38.22	***		1.94		
PBIAS	Political bias	4.13			12.23	**	
L–R	Left–right scale	10.94	*		12.88	**	

Table 2: Dependency of hand-engineered features against the signature count (p and H) and deep hidden features (p_{hidden}). ADD is not applicable for the US government petitions dataset. $p < 0.001$, $p < 0.01$ and $p < 0.05$ are denoted as "***", "**" and "*", respectively.

Overall our proposed approach with the auxiliary loss and hand-engineered features ($\text{CNN}_{\text{regress+ord+feat}}$) provides a reduction in MAE over $\text{CNN}_{\text{regress}}$ by 2.1% and 3.2%, and SVR by 7.2% and 13.7% on the UK and US datasets, resp. Although the ordinal classification performance is not very high, it must be noted that the data is heavily skewed (only 2% of the UK test-set falls in the $[10000, 100000)$ and $[100000, \infty)$ bins put together), and we tuned the hyper-parameters wrt the regression task only.

5 Conclusion and Future Work

This paper has targeted the prediction of the popularity of petitions directed at the UK and US governments. In addition to introducing a novel task and dataset, contributions of our work include: (a) we have shown the utility of an auxiliary ordinal regression objective; and (b) determined which hand-engineered features are complementary to our deep learning model. In the future, we aim to study other factors that can influence petition popularity in conjunction with text, e.g., social media campaigns, news coverage, and early growth rates.

Acknowledgements

We thank the reviewers for their valuable comments. This work was funded in part by the Australian Government Research Training Program Scholarship, and the Australian Research Council.

References

Molly Asher, Cristina Leston Bandeira, and Viktoria Spaiser. 2017. Assessing the effectiveness of e-petitioning through Twitter conversations. *Political Studies Association (UK) Annual Conference*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Zsolt Bitvai and Trevor Cohn. 2015. Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 180–185.

Catherine Bochel and Hugh Bochel. 2017. 'Reaching in'? The potential for e-petitions in local government in the United Kingdom. *Information, Communication & Society*, 20(5):683–699.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Russell J. Dalton, Susan E. Scarrow, and Bruce E. Cain. 2003. *Democracy Transformed?: Expanding Political Opportunities in Advanced Industrial Democracies*. Oxford University Press.

Ahmed Said Elnoshokaty, Shuyuan Deng, and Dong-Heon Kwak. 2016. Success factors of online petitions: Evidence from change.org. In *49th Hawaii International Conference on System Sciences*, pages 1979–1985.

Scott A. Hale, Helen Margetts, and Taha Yasseri. 2013. Petition growth and success rates on the UK No. 10 Downing Street website. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 132–138.

Hansard. 2016. *Audit of Political Engagement 13*. Hansard Society, London, UK.

Shih-Wen Huang, Minhyang Mia Suh, Benjamin Mako Hill, and Gary Hsieh. 2015. How activists are both born and made: An analysis of users on change.org. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 211–220.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Ling Li and Hsuan-Tien Lin. 2007. Ordinal regression by extended binary classification. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 865–872.

Ralf Lindner and Ulrich Riehm. 2011. Broadening participation through e-petitions? An empirical study of petitions to the German parliament. *Policy & Internet*, 3(1):1–23.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alicja Piotrkowicz, Vania Dimitrova, Jahna Otterbacher, and Katja Markert. 2017. Headlines matter: Using headlines to predict the popularity of news articles on Twitter and Facebook. In *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 656–659.

Julia Proskurnia, Przemyslaw Grabowicz, Ryota Kobayashi, Carlos Castillo, Philippe Cudré-Mauroux, and Karl Aberer. 2017. Predicting the success of online petitions leveraging multidimensional time-series. In *Proceedings of the 26th International Conference on World Wide Web*, pages 755–764.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1650–1659.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Systems Research and Behavioral Science*, 7(4):484–498.

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin, and Julian Brooke. 2017. Joint sentence-document model for manifesto text analysis. In *Proceedings of the 15th Annual Workshop of the Australasian Language Technology Association (ALTA)*, pages 25–33.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 175–185.

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.

Taha Yasseri, Scott A Hale, and Helen Z Margetts. 2017. Rapid rise and decay in petition signing. *EPJ Data Science*, 6(1):20.

Chapter 4

Conclusion and Future Work

In this chapter, we discuss how the different contributions of this thesis (as presented in Chapter 3) tie together to achieve an overall goal — to improve political accountability and voter advocacy in democracy. To this end, we first recapitulate the research contributions of this thesis (Section 4.1), and summarise the approaches used in each work for political text analysis. We then discuss how each of these research methods can be used together to address end-to-end political analysis tasks (Section 4.2). Next, we summarise other work that has been published in related domains subsequent to our work being published, which provides the basis for both reflections on the contributions of this work and possible extensions as future work (Section 4.3). Finally, we provide a summary of the thesis.

4.1 Research questions

Our target research goals included: (1) automating party characteristics analysis using election campaign data and tracking progress made by governing party towards its election promises, which can improve political accountability; and (2) promoting voter advocacy — predicting petition popularity, which allows possible intervention by choosing suitable campaign strategies to increase reach among voters.

The finer-level research goals we have targeted over election campaign text can help to dissect the policy proposal for various political analyses. We analyzed the utility of these tasks towards

characterizing political parties in terms of ideology and issue salience. Since labeled data is expensive, and known to be a major bottleneck in political analysis [Cardie and Wilkerson (2008)], we use a semi-supervised learning strategy for each of the tasks.

A primary task in political analysis is to find the policy issues discussed in text. We use manifestos for this task, because of its popularity among political scientists as a canonical source, as well as the availability of digitized content across many countries and over time through the Comparative Manifesto Project (CMP) [Volkens et al. (2011)]. We use the popular label schema proposed by CMP, which has 56 fine-grained classes (Section 3.1) for policy categorization, and acts as an anchor for further analysis. For example, *Will implement the national energy guarantee with a higher emissions reduction target – 45% by 2030* discusses an *energy*-related policy. Many popular analyses in political science use policy issue-coded text through aggregation of code labels, giving rise to characteristic scores such as ideology and issue salience. For instance, if a party manifesto has more pro-*economy* sentences than pro-*environment* ones, then it is said to be right-leaning, as it gives more relative importance to the former than latter policy issue. Typically sentence- and document-level tasks are handled separately, and document-level tasks are addressed based solely on manually annotated sentence-level labels. In this thesis we automated both the tasks, and leveraged the inter-dependencies between the tasks. Specifically, we modeled the dependencies between policy issue classification (sentence-level) and ideology score regression (document-level) using a combined structured loss function. We showed that the joint model improves the performance of both the sentence- and document-level tasks, especially in semi-supervised settings. Though the aggregated ideology scores are relatively easy to compute, and are available for many manifestos, political scientists consider expert survey scores such as Chapel Hill expert survey [Bakker et al. (2015)] as gold-standard data. Because they are not estimated in isolation for each document (unlike the aggregated scores) and are more contextual, i.e., they consider relative and temporal shifts between parties across countries, coalition effects, etc. But expert survey scores are not available for many democracies and election periods, hence we model the contextual dependencies using Probabilistic Soft Logic [Bach et al. (2017)] (PSL), to obtain a re-calibrated ideology score, which we found to be closer to expert survey scores. An important conclusion was that we showed how PSL can be used to model dependencies motivated from domain knowledge to better predict the party characteristics.

Knowing what policy issues are mentioned in each sentence (from the model in Section 3.1) does not convey its function in the broader political discourse. For instance, *We are currently on*

track to exceed 3-degrees of global warming, which would be utterly catastrophic not just for Australia but for all nations around the world conveys the party's belief on climate change, and *Will focus on reducing waste in our rivers and oceans and on reducing the pressures on our environment by increasing our recycling capacity, reducing plastic waste and reducing food waste* discusses their proposed action. Knowing the context or rhetorical function of an utterance is necessary to distinguish its purpose. As a step towards the automatic detection of rhetorical functions, we modeled the *speech acts* of campaign text, by combining traditional classes [Searle (1976); Austin (1962)] with domain specific classes sourced from the political science literature [Eder et al. (2017)]. Since the utterances can be targeted at any of the parties, we also considered the task of *target* party classification. For example, this text from Liberal party, *Labor is planning to impose more than $200bn of new taxes on the Australian economy over the next decade* refers to its opposite party's (Labor) actions. While *the Coalition would also cut taxes for middle and high-income earners from 2024 by flattening brackets so those earning between $45,000 and $200,000 all paid a marginal rate of 30%* talks about its own (Liberal party) proposed action on tax reforms. We use deep learning to model the two tasks, and since obtaining large amounts of labeled data is difficult, we proposed a semi-supervised extension of the approach to handle the low supervision scenario. We use a multi-view learning based semi-supervised technique (teacher-student ensemble), which learns the target model under different configurations of parameters, in order to make the model's predictions more robust. In our manifesto analysis work (Section 3.1), we handled sentence- and document-level tasks together using a multi-task structured loss to capture the inter-dependencies, and showed it can provide gains in performance. In this work, we observed that a simple multi-task objective for speech acts and target party classification tasks did not provide any significant improvements. This calls for a better way to capture the task dependencies by building on other choices of multi-task architectures [Ruder (2017)], and it is more compelling to analyze its effectiveness in combination with the teacher-student model. For this work, we annotated a dataset using speeches and media-releases released by Labor and Liberal party during 2016 Australian Federal election. Because campaign speeches convey both the political function as well as the general persuasive function, they are a natural fit for this study. Having obtained the lower-level rhetorical functions, to understand how well they characterise political parties, we analyzed the distribution of rhetorical functions across different policy issues. We found that speech act usage patterns varied between the two ideologically-opposed parties. Since the dataset contains text only pertaining to two parties during a single election (which gives two data points), we could not perform prediction of party characteristics as we did in our previous task (Section 3.1).

Among messages which convey future promised actions, it is necessary to grade their specificity, e.g., rhetorical vs. concrete. In the case of, *We will stand up for the country* vs. *We will invest a billion dollars for Medicare*. We used a label schema proposed by political scientists [Pomper and Lederman (1980)], which has seven grades of promise specificity, and annotated a dataset based on Australian election manifestos. Since the specificity categories are graded (i.e., one is more specific than another) and the difference in commitment between any two categories is not linear, we proposed an ordinal regression approach for predicting promise text specificity. We once again proposed a semi-supervised extension using the teacher-student framework (which was effective for the speech act classification task, Section 3.2). We empirically showed that the ordinal regression formulation of the task performs better than posing the task as either a classification or regression problem. In terms of capturing downstream party characteristics, we showed that promises distinguished by their specificity, made on various policy issues, can better capture party ideology and issue salience compared to simple frequency counts of policy mentions (Section 3.3). For instance, the idea is to differentiate between discussing *environmental* policies and making concrete promises to improve the environment. For estimating party ideology on a continuous left-right scale, we used a PSL model similar to the manifesto position analysis task (Section 3.1), which incorporates the required domain information as dependencies.

Next we focused on an *election promise tracker* (Section 3.4), which is an explicit way to improve accountability, by keeping track of progress made towards promises given by the winning party. This forms the core of representative democracy. All the work in the political science domain for this particular task has relied on manual annotations, including efforts from news organizations such as the Australian Broadcasting Corporation to keep track of promises for a particular government. We formulated a two-stage task which can help to (semi-)automate this task — to pick track-worthy promises from an election campaign, and track their progress based on textual evidence. We use supervised learning for this task, by curating a dataset from existing resources covering Australia, Canada, the UK, and the US, and evaluated various text representations. We analyzed the patterns learnt by the model to rank track-worthiness, by manually interpreting those instances it predicted wrongly, and found that specificity of promises was an important attribute to be tracked by the news organizations.

Lastly, we worked on a task relating to voter advocacy. All other work discussed above deals with political parties and their role in democracy. Here, we analyze petitions, which are an explicit form of voter advocacy for policy changes (Section 3.5). First, we model the popularity

of petitions using text. This can help to rank future petitions, with potential applications such as boosting the visibility of potentially popular petitions, either by promoting on the first page of a petition site (homepage) or allowing the relevant petition committee to dedicate resources to ones that are more likely to be endorsed by other citizens. Previous work relied on hand-crafted features extracted from text, and used correlation for analysis, which helps to identify textual features that are persuasive, and can potentially help petition committee take actions accordingly to improve the reach of petitions. We proposed a supervised learning approach to model petition popularity as a function of its text, and analyzed what the model captured (and did not) based on a set of hand-crafted features. We used a linear regression model to compute the dependencies between these features and the hidden representation of the deep learning model, and found that the model captures aspects such as syntax and sentiment, but does not capture political bias-related features (though it does capture the target popularity). There are several other confounding factors for this analysis, such as virality of the topic at the particular time say through TV news, whether the petition appeared on the homepage of the website or not, and response of a petition committee to a previous similar petition, that deserve future attention.

In the later two tasks, the major contribution lies in automating the tasks for the political setting. We used supervised learning approaches, and focused on understanding what the model captured, using manual inspection of the output and by correlating the hidden representation against a set of hand-crafted features. We discuss the commonalities among different approaches used in this work in Section 4.1.1, and discuss several immediate extensions possible from our work (in Section 4.1.2).

4.1.1 Approaches

An overview of the technical approaches used in this thesis and their commonalities is discussed here. Among the deep learning methods we used for campaign text analysis, one commonality is the use of semi-supervised learning. We use the popular strategy of multi-task learning when we have multiple dependent tasks, and a teacher-student framework to learn a single task in self-ensembling style, to leverage unlabeled data for learning the models. Combining both the approaches is a potential direction for future work (discussed further in Section 4.1.2). Secondly, we have explored various applications requiring a regression formulation, including ordinal regression. Note that regression tasks are relatively less explored in the field of NLP. We also

investigated and showed the use of an auxiliary ordinal regression task, in addition to the main petition popularity regression task (Section 3.5).

The lower-level text annotations are usually used for analyzing higher-level political characteristics (e.g., issue salience of party). Hence we study how well the output of automatic approaches can capture downstream characteristics, wherever applicable (Sections 3.1, 3.2 and 3.3). Though deep learning models provide large gains in performance, they are very hard to interpret without any additional probes. Being able to interpret predictive models can help to provide insights to social scientists with respect to the problem of interest. To this end, we use simple approaches, which are driven by domain knowledge and are based on manual effort — we used manual inspection of the output for promise track-worthiness estimation (Section 3.4) and modeled the dependency between the (deep) hidden representation and a set of hand-crafted features for petition popularity analysis (Section 3.5). There is a potential scope to extend it with more sophisticated techniques (discussed in Section 4.3).

Another commonality, in addition to employing deep learning approaches to model text, is the use of logic to incorporate domain information in order to build structured models for political analyses. Political science has rich domain information, owing to the manual studies done by social scientists over many years. Hence we should ideally look to incorporate such available knowledge into the predictive models. To enable this, logic-based representations like first-order logic, are powerful in capturing domain information in the form of unary predicates, negation, relations using binary predicates, and quantifiers. Since many cases require handling uncertainty for reasoning, the combination of first-order logic and probabilistic graphical models are more popular [Richardson and Domingos (2006)]. In this thesis, we use one such approach called probabilistic soft logic [Bach et al. (2017)] (PSL), which uses weighted logic to encode dependencies over random variables, which specifies a ground Markov random field. We used PSL for predicting party left-right ideology position (Section 3.1 and 3.3). Existing work which employs PSL for political stance prediction has used speaker party affiliation to capture relational information, in addition to the word usage [Andrzejewski et al. (2011); Johnson et al. (2017); Johnson and Goldwasser (2018)]. In our work we exploited party authorship of campaign text, coalition information, and temporal dependencies for ideology prediction, which have been shown to influence party position by political scientists [Greene (2016)].

4.1.2 Possible immediate extensions

In social science tasks, the lower-level NLP outputs are used to drive higher-level political analyses. Similarly, in this thesis, for ideology prediction based on policy mentions and promise specificity (Section 3.1 and 3.3), we used the proposed deep learning models to obtain sentence-level output, and then the aggregated scores were used in the PSL model to predict the ideology positions. A possible extension is to build a joint model, which can provide both the basic text unit predictions and the required higher-level ideology scores which the political scientists care about, with suitable encoding of domain-driven dependencies. This model can help to back-propagate the influence of domain information (in the form of related dependencies) onto the lower-level tasks.

We used a teacher-student framework based semi-supervised learning for speech acts classification and promise specificity prediction, and can apply a similar framework to other tasks — including policy issue classification, petition popularity prediction and promise progress tracking — to handle sparse supervision. Especially, to study whether the multi-task model for policy analysis benefits from using a teacher-student network extension, can be a novel empirical analysis. Liu et al. (2019) have shown that multi-task teacher-student networks improve the efficiency of knowledge distillation for pretrained models. Notably, all our work assumed some level of supervision, and it is a worthy challenge to model the various tasks with transfer learning, or distant supervision. Example scenarios are, for speech acts, policy topic and promise specificity categorization tasks, where there are annotated datasets from politics or other related domains, at least with partial overlap over target categories, which can be used to bootstrap the models accordingly. For instance, in Ko et al. (2019), the authors proposed a transfer learning approach to predict social media (e.g., Twitter, Yelp) text specificity using news domain specificity annotations. Another scenario in relation to policy theme classification is to use the set of keywords provided by domain experts (for instance, in the Comparative Manifesto and Comparative Agendas Projects) to obtain distantly labeled examples.

For campaign text analysis, the goal of speech act classification is to identify rhetorical functions such as what is the belief of a party, what actions they propose to take that build on their beliefs, and how they compare with other parties' actions. Such discourse connections can be formed without any additional steps for media-release documents, because they focus on a single issue. But speech transcripts and manifestos cover a wide variety of issues, and hence require segmentation into issues, then analysis of rhetorical functions that can provide a coherent

discourse analysis. Further, multiple speeches and media-releases are made during an election cycle, hence to study the evolution of a policy proposal, it first requires aligning different text based on thematic relations and then analyzing policy proposals made in one text conditioned on the relevant previous announcements, to establish inter-document discourse connections.

We have only scratched the surface of the promise tracking problem (Section 3.4) — defining a new task and label schema, curating existing resources to build a dataset, and evaluating contemporary approaches to serve as a baseline for future work. We can extend it in many different directions. First, we posed promise status prediction as a three-way classification problem, which is a simplification of the actual task, which has several grades of progress. An immediate extension of this work can be to handle other status grades such as *stalled* and *partially fulfilled*. Second, the natural language inference problem is non-trivial because it requires understanding numbers, ratios, etc. More work needs to be done to assess whether the model understands quantities of different types, and use a more powerful model if required. Thirdly, as in the case of fact-checking [Thorne et al. (2018)] we can develop automatic ways of creating large amounts of labeled data, to enable training of large, sophisticated models. Given a broad range of media sources, it will be necessary to perform a critical analysis of the suitability of each source and individual articles to track promise progress. As both the news media and individual news articles are shown to have different ranges of political biases [Baly et al. (2020)], it becomes very important to rank them based on objectivity [Lex et al. (2010)] in the automated analysis system.

In this work we used different slices of data for analysis, with some level of overlap for the Australian setting. This allows us to answer further inter-related questions by modeling logically inter-connected datasets, for example, to analyze policy issues and discourse, and the interaction between parties, promises made, promises kept or broken, and voter advocacy for a given set of countries in the same period of time. This can help to achieve a wide range of analyses, some of which are discussed in Section 4.2.

4.2 End-to-end analysis

Here we discuss how the different components can be put together to build an end-to-end application that can provide detailed political insights. In this thesis, we used different data for each task, but here propose a combined approach assuming the individual tasks are well aligned in

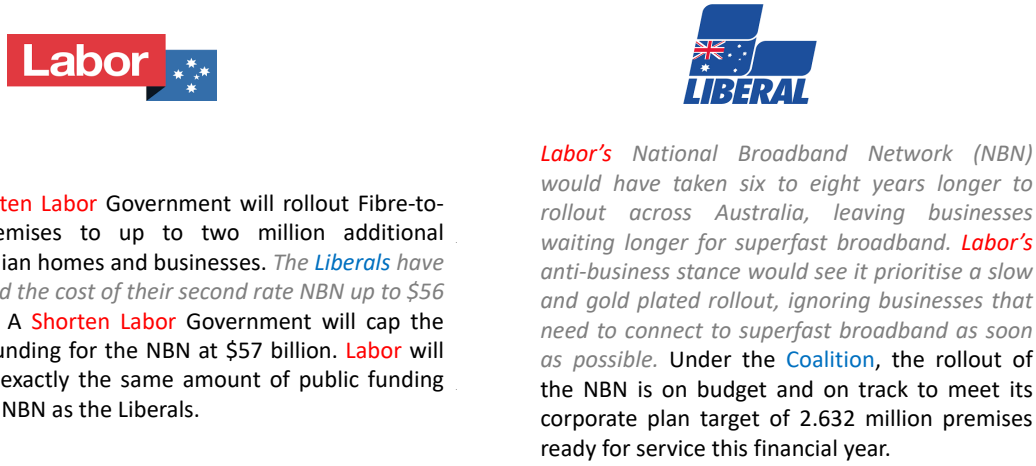


FIGURE 4.1: An example from 2016 Australian election — interaction between Labor and Liberal parties through press-releases. It shows the target party in color, and criticism of other party's policy (interaction) in italics, and the promises made by parties towards broadband communication.

terms of country, time and data coverage to achieve holistic political analysis. We provide two different views of analysis: (a) dissecting election campaign data to provide easy-to-consume summaries for voters, and (b) assessing policy implementation in terms of voter expectations. The idea would be to align election promises, which set the expectations of voters with respect to policies, and consider voter advocacy as both feedback on existing policies and a dynamic way of engaging voters in policy agenda setting. This thesis is more concentrated on (a), and touches only lightly on (b). But it should be noted that there is potentially much more work that can be done towards both directions.

(a) **Dissecting campaign data can provide an easy-to-consume summary of vast amounts of text**. Some useful analyses that can be achieved with this processing are discussed as follows.

How party positions and interactions between parties evolve over time on various policy issues provide valuable insights into policy formation (see Figure 4.1 for a static snapshot). We can recover interaction patterns, such as how questions and criticisms are conveyed, and whether a party asks for updates, passes condemnatory messages, or seeks factual responses [Zhang et al. (2017)]. Also of importance is, how they are responded to — such as denial, justification, counter-criticism and providing excuses [Naderi and Hirst (2017, 2018)]. Zhang et al. (2017) and Naderi and Hirst (2017, 2018) used parliamentary questions and responses for their studies, which are already aligned. But with campaign text, explicit alignment of messages between parties is required on the basis of policy issues, which can be achieved by building on the approach proposed by Menini and Tonelli (2016). Across elections, parties tend to change their

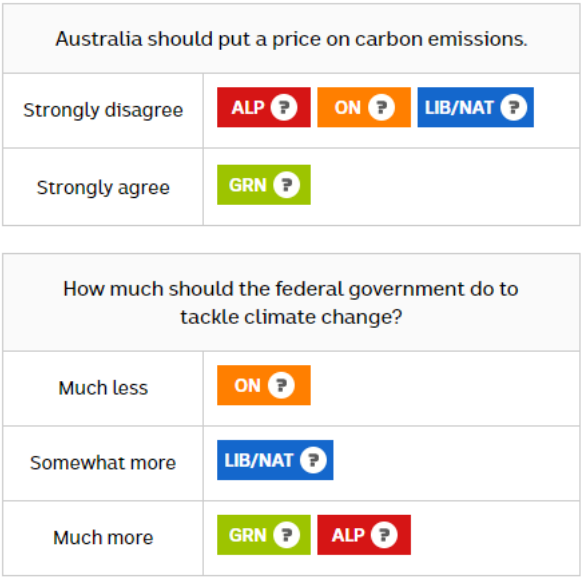
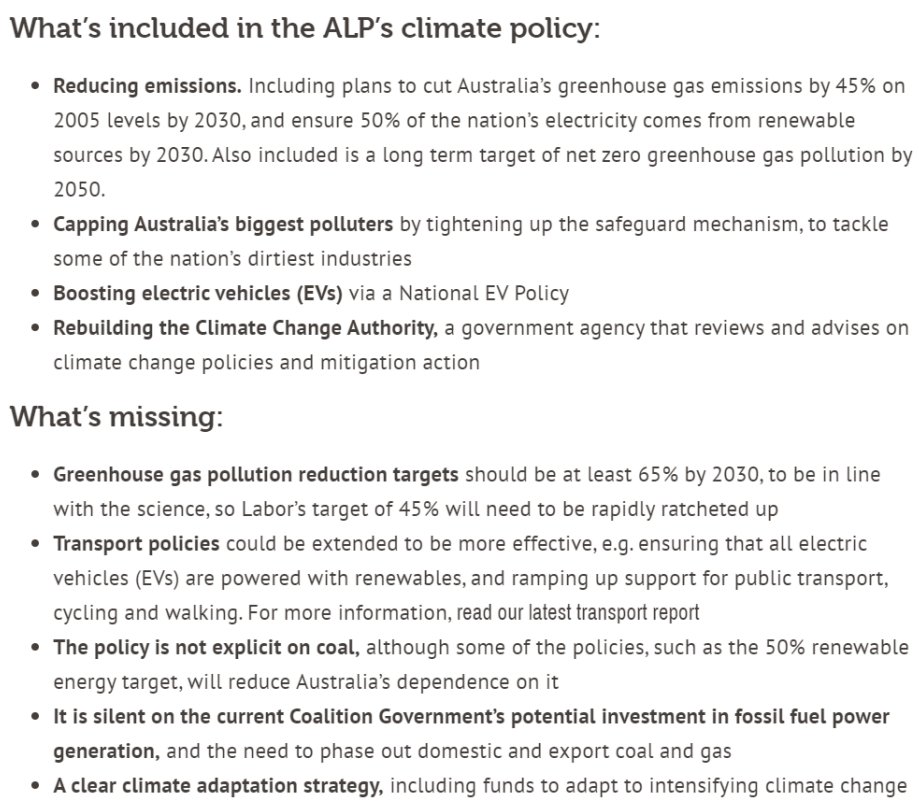


FIGURE 4.2: Snapshot taken from Vote Compass: Analysis of Australian political parties' (LIB=Liberals, NAT=Nationals, ALP=Labor Party, GRN=Greens, ON=One Nation) position

positions based on voter support shifts and electoral performances [Schumacher et al. (2013)]. For example, in the US 2020 Presidential elections, Michael Bloomberg apologized for expanding "stop and frisk" policing during his tenure as the Mayor of New York City, which is seen as discriminatory against people of color.[1] Hence it is useful to analyze party or candidate positions over time to automatically find whether there is a shift, and it can also help to verify the claims involved in inter-party interactions. For example, in the 2008 US Presidential elections, Democratic candidate Obama made this statement on withdrawing troops from Iraq: *I am going to do a thorough assessment when I'm there. I'm sure I'll have more information and continue to refine my policy*. This was targeted by the Republican side as: *Since announcing his campaign... the central premise of Barack Obama's candidacy was his commitment to begin withdrawing American troops from Iraq immediately, but, today, Barack Obama reversed that position proving once again that his words do not matter*. But, by analyzing Obama's position over time, PolitiFact found that he did not reverse it.[2]

Other than analyzing individual party positions, it is very useful to compare policy positions across parties to build a micro-summary. This is motivated by projects such as VoteCompass,[3] which analyzes party views; see Figure 4.2 for an example. It should also be noted that the example from VoteCompass contains graded stance, and not the traditional binary stance [Küçük

[1] https://www.usnews.com/news/elections/articles/2019-11-29/bloomberg-biden-warren-and-the-flip-flops-of-the-2020-presidential-election
[2] https://www.politifact.com/article/2008/jul/10/baracks-iraq-flip-flop-nope/
[3] https://votecompass.abc.net.au/



What's included in the ALP's climate policy:

- **Reducing emissions.** Including plans to cut Australia's greenhouse gas emissions by 45% on 2005 levels by 2030, and ensure 50% of the nation's electricity comes from renewable sources by 2030. Also included is a long term target of net zero greenhouse gas pollution by 2050.
- **Capping Australia's biggest polluters** by tightening up the safeguard mechanism, to tackle some of the nation's dirtiest industries
- **Boosting electric vehicles (EVs)** via a National EV Policy
- **Rebuilding the Climate Change Authority,** a government agency that reviews and advises on climate change policies and mitigation action

What's missing:

- **Greenhouse gas pollution reduction targets** should be at least 65% by 2030, to be in line with the science, so Labor's target of 45% will need to be rapidly ratcheted up
- **Transport policies** could be extended to be more effective, e.g. ensuring that all electric vehicles (EVs) are powered with renewables, and ramping up support for public transport, cycling and walking. For more information, read our latest transport report
- **The policy is not explicit on coal,** although some of the policies, such as the 50% renewable energy target, will reduce Australia's dependence on it
- **It is silent on the current Coalition Government's potential investment in fossil fuel power generation,** and the need to phase out domestic and export coal and gas
- **A clear climate adaptation strategy,** including funds to adapt to intensifying climate change

FIGURE 4.3: Snapshot taken from http://climatecouncil.org.au which analyses Australian Labor Party's climate policy.

and Can (2020)], which is a challenge in itself. Though our promises analysis work (Section 3.3) can be used to compare party goals, we do not handle stance classification explicitly, and this is a potential task for future work. Furthermore, specific promises may not always contain all the necessary goals to reach the desired target. For instance, the Climate Council analyzed climate change related policy goals from the Australian Labor Party, and identified gaps based on world knowledge (see Figure 4.3). Identifying gaps in party goals based on the recommendations from different policy research groups or UN goals requires document-level (or multi-document) natural language inference. Galsurkar et al. (2018) attempted the preliminary step of aligning national policy goals with UN sustainability goals, posing it as a search task.

(b) **Aligning election promises with policies and related voter advocacy for change**

A core part of representative democracy is the promises made by political parties, which are guarantees given to voters, to implement them in the form of policies after forming government. A healthy democracy should encourage voter participation in the policy-making process. Assessing these end-to-end requires analyzing data from multiple sources, and aligning them to get a complete view. For example, policy implementation related to Brexit and voter advocacy in

Theresa May has signed the letter that will formally begin the UK's departure from the European Union. Giving official notice under Article 50 of the Lisbon Treaty, it will be delivered to European Council president Donald Tusk later.

Mar 2017

Brexit is a major issue at the UK general election - the reason for calling a general election was to strengthen her hand in negotiations with the EU.

Jun 2017

Revoke Article 50 and remain in the EU.
The government repeatedly claims exiting the EU is 'the will of the people'. We need to put a stop to this claim by proving the strength of public support now, for remaining in the EU. A People's Vote may not happen - so vote now.

Feb 2019

Revoking Article 50 would break the promises made by government to the British people, disrespect the clear instruction from a democratic vote, and in turn, reduce confidence in our democracy

Mar 2019

FIGURE 4.4: An example for voters' advocacy in the UK politics wrt Brexit policy. It shows four events (top to bottom) — Theresa May invoking Article 50; 2017 elections where Conservatives promised Brexit, won and formed the government; voters' petition to revoke Article 50 fetched lots of signatures; Government responds that Article 50 cannot be revoked as it breaks their election promise. All the text descriptions are taken from BBC.

UK politics is given in Figure 4.4. In the 2016 Brexit referendum, UK citizens voted in favor of Brexit.[4] Following that, Theresa May, then Prime Minister of UK, invoked Article 50 in 2017. Then an election was called, where Brexit was a major issue, and the Conservatives set out a 12 point plan for Brexit.[5] After the elections, Conservatives formed government, and there was a voter-led campaign in the form of petitions to revoke Article 50. The petition (in different forms) got a huge number of signatures[6] and thereby attracted the attention of the government, following which there was a debate. The ultimate response from the government was that it could not revoke Article 50 because that would break its promise to the voters.[7] This particular policy change involved voter advocacy in different ways — referendum, election voting, and lastly through petitions. Such dynamic engagement between voters and the government, which indicates a functioning democracy, can be uncovered by jointly analyzing several sources of text such as those explored in this thesis. The joint analysis requires a non-trivial alignment of text from multiple sources, which is beyond the focused contribution of this thesis. Some of the alignment steps required to achieve the end-to-end analysis mentioned in this section are — concrete election promises linked to government actions can help to keep parties accountable and

[4] https://www.bbc.com/news/uk-politics-39422353
[5] https://www.bbc.com/news/uk-politics-39665835
[6] https://www.bbc.com/news/uk-politics-47665929
[7] https://www.bbc.com/news/uk-47711206

voters informed, based on which they can provide feedback; and the policy actions linked to petition issues can help to relate them to people's expectations and identify gaps. In this thesis, we touched upon the challenge of aligning promises to relevant web documents discussing policy updates, by including *relevance* as an additional class and evaluated with a balanced dataset for *relevant* and *irrelevant* categories (Section 3.4). In real-world settings over hundreds of news sources, however, an information retrieval component is required to first identify candidates from a vast set of web documents [Yang et al. (1999)], in addition to the linkage model which can classify whether a document contains any policy-related action relevant to a given promise statement. Further, to link grassroots petition issues to policy actions, there is a need to identify whether a petition calls for action related to gaps in existing policies or the development of new policies, which is a challenge in itself. In the government petition platforms (e.g., UK Government petitions), alignment of petitions to parliamentary proceedings is performed manually,[8] which can potentially be (semi-)automated to reduce cost. The end-to-end alignment tasks can be addressed as a document-level event co-reference resolution problem [Kenyon-Dean et al. (2018)], where existing benchmarks are typically small in size as labeling large amounts of data is expensive, and hence focus must be given to methods which work with minimal supervision.

4.3 Reflections and future work

Here we provide reflections on the thesis contributions and discuss future work, based on recent developments in related areas which can provide additional resources and ideas for both improving our work and developing further ideas. In our work, we used word embeddings built using bag-of-words style training for manifesto policy topic analysis (Section 3.1) and petition popularity prediction (Section 3.5). We used contextual word embeddings for speech act classification (ELMo), promise specificity prediction (ELMo), and promise tracker (BERT, ELMo, USE) (Sections 3.2, 3.3, and 3.4 respectively, reflecting the state-of-art encoding method at the time of the respective publications). In this thesis, we targeted policy position analysis for manifestos written in different languages, by projecting mono-lingual word embeddings to a common space (Section 3.1). There are several more recent papers proposing new and better multi-lingual contextual embeddings such as multi-lingual BERT (M-BERT) [Devlin et al. (2019)] and cross-lingual language models (XLMs) [Conneau and Lample (2019)]. M-BERT is a single language model pre-trained from monolingual corpora in 104 languages. XLMs allow pre-training BERT

[8] https://commonslibrary.parliament.uk/research-briefings/sn06450/

for multiple languages, both using a self-supervised technique that only uses monolingual data, and a supervised approach that leverages parallel data with a cross-lingual language model objective. With the rising popularity of language model based representations, it is an obvious next step to evaluate policy position analysis and petition popularity prediction tasks using contextual embeddings, with multi-lingual policy analysis being a particularly attractive target for cross-lingual language models. Other than policy analysis, all other work dealt with English text only. It will be an important contribution to both the NLP and political science communities to extend all the tasks to a multi-lingual setting.

In all our work, we use pre-trained embeddings for political text analysis, and fine-tune the embeddings based on the end-task objective. On the other hand, Glavaš and Vulić (2018) adapt word vectors to the target domain by fine-tuning them using external knowledge in the form of attract and repel word pairs, and show gains on downstream applications. Sarma et al. (2018) propose a method to combine generic embeddings with domain specific embeddings, which they show to perform better than using both generic and domain specific embeddings (when trained on a smaller target corpus). For language modeling, Soares et al. (2019) propose a BERT-based model for relation learning, and show that fine-tuning BERT by matching the blanks using a relation extraction dataset attains good performance, especially in a few-shot learning setting. The dataset is marked with entities, and for fine-tuning BERT, entities are replaced with a special [BLANK] token. This shows that task-specific fine-tuning strategies can improve on standard techniques in sparse supervision scenarios. Howard and Ruder (2018) propose a approach for fine-tuning (pre-trained) language models, which handles transfer learning and leverages unlabeled data. They use discriminative fine-tuning to tune each layer with different learning rates, and slanted triangular learning rates to achieve faster convergence. Though we use unlabeled data in our semi-supervised learning approaches, we do not evaluate using large amounts of unlabeled text, especially when fine-tuning language models. We believe it is possible to further improve empirical performance, reducing the gap between manual analysis and automatic approaches, with the use of suitable domain adaptation techniques to fine-tune pre-trained distributed text embeddings, both for our target domain and specific tasks.

Secondly, in our teacher-student network-based semi-supervised settings, we use word-level dropout to encourage the learning of diverse networks for self-ensembling. Though word dropout is an effective approach, there is scope to evaluate and even identify custom data augmentation strategies for robust training. Some recent work which explore that direction is discussed here. Rather than using a single data point for augmentation, mixup [Zhang et al. (2018)] performs

interpolation of data pairs to achieve augmentation. Virtual adversarial training [Miyato et al. (2018)] learns additive perturbation to apply to the input which maximizes the likelihood of the output class distribution. Xie et al. (2019) use round-trip translation for data augmentation, which translates text from one language into another language, and then translates it back into the original language to obtain an augmented example. We can use these strategies to achieve task-specific data augmentation for semi-supervised learning. For instance, in the promise specificity prediction task, tokens related to policy details should not be dropped. On the other hand, it is not always the case that a *right-leaning* party will make specific promises on *economy*, hence masking party references could help in learning a party-agnostic representation, and thereby avoid overfitting.

For campaign text analysis, there are many other survey datasets released and used by political scientists — Comparative Candidates Survey (CCS: [CCS (2018)]), Parliaments and Governments Database (ParlGov: [Döring and Manow (2018)]) and Party Facts (PartyFacts: [Döring and Regel (2019)]). CCS contains candidate surveys covering general political involvement of candidates, campaign activities, issue positions of elites, attitudes towards democracy and representation, as well as the personal background of candidates running for election. ParlGov contains information about parties, elections and cabinets for all EU and most OECD democracies (37 countries), from 1945 until recent elections. PartyFacts covers over 5000 parties spread across more than 210 countries, and links with party information from other social science data sets. This constitutes about 15,000 party observations[9] — mass surveys, data handbooks, and various datasets on election results, voting records, party characteristics and party positions. We used ParlGov to obtain the coalition information of political parties (Sections 3.1 and 3.3), that we modeled as a relational dependency for predicting their ideology shifts at each election. But there is more valuable information in ParlGov and other datasets which we can use as features to better characterise parties or use as target party indicators as part of our model. For instance, we can use cabinet details to study each party's policy-issue salience, which can be used to weight their positions. Similar to our work with political parties, we can model individual candidates' positions using their speeches, say on social media, and use CCS indicators as the ground-truth.

There are other threads of political science work that consider other aspects in their analysis, such as how party pledges translate into coalition agreements [Vodová (2020)]. Müller (2020) investigate the media coverage of election promises. Duval and Pétry (2019) study the effect of time (how far a government is through in its mandate) on pledge fulfilment. Duval (2019)

[9] https://partyfacts.herokuapp.com/data/

investigate the media coverage of the fulfillment of electoral pledges, finding that news media are effective at alerting citizens when a pledge is broken. Matthieß (2020) analyze data from 69 elections in 14 countries and show that an incumbent party's electoral outcome is affected by its performance related to previous pledges. Compared to the traditional use of expert evaluation for pledge fulfillment rating, Duval and Pétry (2018) examine citizen evaluation of fulfillment. The authors found that citizen evaluations depend on group identities and a priori beliefs, including partisanship and political trust, which do not increase the accuracy of pledge evaluations. This shows that there are several other social science challenges, related and complementing this thesis, that can be addressed further. Some such challenges include: automatically grading media bias [Fan et al. (2019)] in terms of tracking and reporting a political parties' fulfillment of electoral pledges; analyzing citizens' awareness of party pledges and their view of government performance using social media data; and monitoring voter intent of action towards their satisfaction or dissatisfaction of party performance.

Interpreting what a deep learning model learns is essential to extend the knowledge of social scientists. We performed such analysis for both petition popularity and promise track-worthiness estimation models (Sections 3.5 and 3.4 respectively). Chawla et al. (2019) used our regression approach proposed in Section 3.5 to analyze information captured by hidden representation for an affect prediction task. There are other ways proposed in the literature to interpret deep learning models. Saliency maps can explain a model's prediction by using gradient-based methods to determine the importance of input tokens, e.g., in the form of the gradient of the loss with respect to the tokens [Simonyan et al. (2013)], integrating the gradient along the path to the input [Sundararajan et al. (2017)], or the average gradient over many noisy versions of the input [Smilkov et al. (2017)]. Other gradient based approaches include HotFlip [Ebrahimi et al. (2018)], which performs word-level substitutions in order to change the model's prediction, and Feng et al. (2018) perform input reduction such that it does not change the model's prediction. Wallace et al. (2019a) provide a framework (AllenNLP Interpret) for interpreting NLP models. The toolkit provides several primitives for interpretation (e.g., input gradients), a suite of interpretation methods, and a library of front-end visualization components. Our interpretation strategies rely on manual efforts — hand-crafting intuitive features to study their dependencies against deep hidden representations for the petition popularity prediction task, and manually analyzing predictions for the track-worthiness estimation model. We can adapt these approaches to interpret the model, to grade the importance of the input text directly, for example, what parts of a promise text are important for it to be specific and what portions of a petition text contribute

to its popularity. A major advantage with these approaches is that it is not necessary to exhaustively enumerate all the features manually to interpret models. But, extending these approaches to take domain-knowledge driven features into account can make them even more interpretable.

There are various directions to explore further, towards the goal of improving voter participation through petition platforms. Hale et al. (2018) analyzed petition platform design to improve citizen participation. They showed that the trending petition feature on the homepage has an effect on the distribution of signatures across petitions. Clark et al. (2018) and Vidgen and Yasseri (2020) investigated the influence of geographic features towards petition popularity, finding that support for issues varies across regions, urban versus rural, and specific issues receive support from regions across the country (e.g., Law & Order, work & pay). Vidgen and Yasseri (2020) also found that petition popularity can be influenced by external events, such as the Brexit referendum. Shen et al. (2019) proposed a method that uses hand-crafted features to guide the learning of a model, by explicitly attending to the feature indicators. They evaluated the approach using UK government petition dataset, and showed improvement over simple concatenation to augment deep and hand-crafted features, which we used in our work (Section 3.5). From these approaches, it is evident that incorporating more attributes is necessary to avoid confounds in our analysis, including a petition's rank on the website homepage, popularity of the issue (say its coverage on TV news), and target region of a petition's proposed action (e.g., state versus nation-wide). Also it is important to use alternate modeling approaches that can better leverage domain information to improve predictive accuracy. We can not only use observed domain information but also unobserved (latent) attributes by suitably modeling the text, especially for analysis over time and across countries [Landeiro et al. (2019)].

4.4 Summary

In this thesis we have focused on automating political science tasks using NLP. Our primary focus was on modeling election campaign text across different dimensions to enable various analyses. We also considered promise progress tracking, which we believe is a seed for a thread of future extensions. Lastly, we addressed voter advocacy, through modelling petition popularity. We believe the work done here and the resources created can support future work on these topics. We have proposed new tasks for the NLP community, including promise specificity prediction and promise progress tracking. We developed modeling solutions for each of the tasks based on the task constraints and available data. Our observations from this work are

that political science is a rich domain for the NLP community, as it contains a wide range of data types and challenging problems where NLP has a lot to offer, rooted in the nature of political communication, which contains a variety of modalities of expression; defining tasks and developing needed datasets requires extensive domain knowledge; in many cases the available data might not be sufficient to build automated approaches, for example, expert surveys for a particular party in a country over 20 years might only cover 5 elections, leaving few instances for learning; and lastly, any assessment of party performance for accountability or voter advocacy can be influenced by many confounding factors.

We hope our research will continue to drive the field forward, serving as a stepping stone for future research on political text analysis, especially on challenges related to analysing canonical sources of text.

Bibliography

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

Gavin Abercrombie and Riza Theresa Batista-Navarro. 'Aye'or 'No'? speech-level sentiment analysis of Hansard UK parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 4173–4180, 2018a.

Gavin Abercrombie and Riza Theresa Batista-Navarro. Identifying opinion-topics and polarity of parliamentary debate motions. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis (WASSA)*, pages 280–285, 2018b.

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

Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery (LinkKDD)*, pages 36–43, 2005.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 1–9, 2011.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1171–1177, 2011.

Pablo Aragón, Diego Sáez-Trumper, Miriam Redi, Scott Hale, Vicenç Gómez, and Andreas Kaltenbrunner. Online petitioning through data exploration and what we found there: A dataset of petitions from Avaaz.org. In *Twelfth International AAAI Conference on Web and Social Media (ICWSM)*, pages 474–480, 2018.

Molly Asher, Cristina Leston Bandeira, and Viktoria Spaiser. Assessing the effectiveness of e-petitioning through Twitter conversations. *Political Studies Association (UK) Annual Conference*, 2017.

S. Ashiagbor. *Political Parties and Democracy in Theoretical and Practical Perspectives: Developing Party Policy*. National Democratic Institute for International Affairs, 2013.

Pepa Atanasova, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 1: Check-worthiness. In *CEUR Workshop Proceedings, CEURWS. org*, 2018.

J. L. Austin. *How to do things with words*. Clarendon Press, Oxford, 1962.

Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss Markov random fields and probabilistic soft logic. *The Journal of Machine Learning Research*, 18(1):3846–3912, 2017.

Patrik F Bakken, Terje A Bratlie, Cristina Sánchez-Marco, and Jon Atle Gulla. Political news sentiment analysis for under-resourced languages. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING): Technical Papers*, pages 2989–2996, 2016.

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

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media (LASM)*, pages 49–58, 2013.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.404.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 2: Factuality. In *CEUR Workshop Proceedings, CEURWS. org*, 2018.

Frank R Baumgartner, Christoffer Green-Pedersen, and Bryan D Jones. Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7):959–974, 2006.

Kenneth Benoit and Michael Laver. *Party policy in modern democracies*. Routledge, 2006.

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

William L Benoit. The functional approach to presidential television spots: Acclaiming, attacking, defending 1952–2000. *Communication Studies*, 52(2):109–126, 2001.

William L Benoit. Topic of presidential campaign discourse and election outcome. *Western Journal of Communication (includes Communication Reports)*, 67(1):97–112, 2003.

Adam Bermingham and Alan Smeaton. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, 2011.

Shaun Bevan. Gone fishing: The creation of the Comparative Agendas Project master codebook. In *Comparative Policy Agendas: Theory, Tools, Data*. Oxford University Press, Oxford, 2019.

Sumit Bhatia and P Deepak. Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 79–84, 2018.

Felix Biessmann. Automating political bias prediction. In *arXiv:1608.02195*, 2016.

Aritz Bilbao-Jayo and Aitor Almeida. Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, 14(11):1–11, 2018.

Knud Böhle and Ulrich Riehm. E-petition systems and political participation: About institutional challenges and democratic opportunities. *First Monday*, 18(7), 2013.

Adam Bonica. Database on Ideology, Money in Politics, and Elections: Version 3.0. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/O5PX0B, 2019.

Camille Borrett and Moritz Laurer. The CEPS EurLex dataset: 142.036 EU laws from 1952-2019 with full text and 23 variables, 2020. URL https://doi.org/10.7910/DVN/0EGYWY.

John Brandt. Text mining policy: Classifying forest and landscape restoration policy agenda with neural information retrieval. *Forestry*, 2(5):3, 2019.

Bastiaan Bruinsma and Kostas Gemenis. Validating wordscores: The promises and pitfalls of computational text scaling. *Communication Methods and Measures*, 13(3):212–227, 2019.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 1506–1515, 2011.

Clint Burford, Steven Bird, and Timothy Baldwin. Collective document classification with implicit inter-document semantic relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 106–116, 2015.

Sarah Cameron and Ian McAllister. Australian Election Study, 1987-2019 Trends, 2019. URL http://dx.doi.org/10.26193/231VJS.

Nick Campbell. On the use of nonverbal speech sounds in human communication. In Anna Esposito, Marcos Faundez-Zanuy, Eric Keller, and Maria Marinaro, editors, *Verbal and nonverbal communication behaviours*, pages 117–128. Springer, 2007.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 438–444, 2015.

Claire Cardie and John Wilkerson. Text annotation for political science research. *Journal of Information Technology & Politics*, 5(1):1–6, 2008.

CCS. CCS 2018. Comparative Candidates Survey Module II – 2013–2018 (Dataset - Cumulative File). Distributed by FORS, Lausanne, 2018.

Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets. *Engineering Applications of Artificial Intelligence*, 51:50–57, 2016.

Kushal Chawla, Sopan Khosla, and Niyati Chhaya. Gated convolutional encoder-decoder for semi-supervised affect prediction. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 237–250, 2019.

Christopher Andreas Clark and Santosh Divvala. Looking beyond text: Extracting figures, tables and captions from computer science papers. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas: AAAI Workshop*, 2015.

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

Stephen Clark, Nik Lomax, and Michelle A Morris. Classification of Westminster Parliamentary constituencies using e-petition data. *EPJ Data Science*, 6(1):16, 2017.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7057–7067, 2019.

Russell J. Dalton, Susan E. Scarrow, and Bruce E. Cain. *Democracy Transformed?: Expanding Political Opportunities in Advanced Industrial Democracies*. Oxford University Press, 2003.

David F Damore. The dynamics of issue ownership in presidential campaigns. *Political Research Quarterly*, 57(3):391–397, 2004.

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

Christopher J Deering and Forrest Maltzman. The politics of executive orders: Legislative constraints on presidential power. *Political Research Quarterly*, 52(4):767–783, 1999.

Shrey Desai, Barea Sinno, Alex Rosenfeld, and Junyi Jessy Li. Adaptive ensembling: Unsupervised domain adaptation for political document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4712–4724, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1195–1198, 2010.

Holger Döring and Philip Manow. Parliaments and governments database: Information on parties, elections and cabinets in modern democracies. *Development version, accessed*, 10, 2018.

Holger Döring and Sven Regel. Party facts: A database of political parties worldwide. *Party Politics*, 25(2):97–109, 2019.

Keith Dowding, Andrew Hindmoor, Richard Iles, and Peter John. Policy agendas in Australian politics: The governor-general's speeches, 1945–2008. *Australian Journal of Political Science*, 45(4):533–557, 2010.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2368–2378, 2019.

Rory Duthie and Katarzyna Budzynska. A deep modular RNN approach for ethos mining. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4041–4047, 2018.

Dominic Duval. Ringing the alarm: The media coverage of the fulfillment of electoral pledges. *Electoral Studies*, 60:102041, 2019.

Dominic Duval and François Pétry. Citizens' evaluations of campaign pledge fulfillment in Canada. *Party Politics*, 2018.

Dominic Duval and François Pétry. Time and the fulfillment of election pledges. *Political Studies*, 67(1):207–223, 2019.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 31–36, 2018.

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

Ahmed Said Elnoshokaty, Shuyuan Deng, and Dong-Heon Kwak. Success factors of online petitions: Evidence from Change.org. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1979–1985, 2016.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6344–6350, 2019.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3719–3728, 2018.

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

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

Jonathan Galsurkar, Moninder Singh, Lingfei Wu, Aditya Vempaty, Mikhail Sushkov, Devika Iyer, Serge Kapto, and Kush R Varshney. Assessing national development plans for alignment with sustainable development goals via semantic search. In *Proceedings of The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 7753–7758, 2018.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, 2019.

Manish Gaurav, Amit Srivastava, Anoop Kumar, and Scott Miller. Leveraging candidate popularity on Twitter to predict election outcome. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNA-KDD)*, pages 1–8, 2013.

Clément Gautrais, Peggy Cellier, René Quiniou, and Alexandre Termier. Topic signatures in political campaign speeches. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2342–2347, 2017.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Fabrizio Gilardi and Bruno Wüest. Text-as-data methods for comparative policy analysis. *Retrieved from* https://www.fabriziogilardi.org/resources/papers/Gilardi-Wueest-TextAsData-Policy-Analysis.pdf, 2018.

Jon Gillick and David Bamman. Please clap: Modeling applause in campaign speeches. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 92–102, 2018.

Goran Glavaš and Ivan Vulić. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-HLT)*, pages 34–45, 2018.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 42–46, 2017a.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 688–693, 2017b.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Computational analysis of political texts: Bridging research efforts across communities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL): Tutorial*, pages 18–23, 2019.

Christoffer Green-Pedersen. Bringing parties into parliament: The development of parliamentary activities in Western Europe. *Party Politics*, 16(3):347–369, 2010.

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

Justin Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

Loni Hagen, Teresa M Harrison, Özlem Uzuner, Tim Fake, Dan Lamanna, and Christopher Kotfila. Introducing textual analysis tools for policy informatics: A case study of e-petitions. In *Proceedings of the 16th Annual International Conference on Digital Government Research (DG.O)*, pages 10–19, 2015.

Scott A. Hale, Helen Margetts, and Taha Yasseri. Petition growth and success rates on the UK No. 10 Downing Street website. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci)*, pages 132–138, 2013.

Scott A Hale, Peter John, Helen Margetts, and Taha Yasseri. How digital design shapes political participation: A natural experiment with social information. *PLOS ONE*, 13(4):1–20, 2018.

Andrew Halterman. Geolocating political events in text. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 29–39, 2019.

Hansard. *Audit of Political Engagement 13*. Hansard Society, London, UK, 2016.

Kazi Saidul Hasan and Vincent Ng. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 816–821, 2013.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM)*, pages 1835–1838. ACM, 2015.

Libby Hemphill and Andrew J Roback. Tweet acts: how constituents lobby congress via Twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 1200–1210, 2014.

John Heritage and David Greatbatch. Generating applause: A study of rhetoric and response at party political conferences. *American Journal of Sociology*, 92(1):110–157, 1986.

Frederik Hjorth, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics*, 2(2):1–9, 2015.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 328–339, 2018.

Michael Howlett. Policy analytical capacity and evidence-based policy-making: Lessons from canada. *Canadian Public Administration*, 52(2):153–175, 2009.

Shih-Wen Huang, Minhyang Mia Suh, Benjamin Mako Hill, and Gary Hsieh. How activists are both born and made: An analysis of users on Change.org. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 211–220, 2015.

Pekka Isotalus. Analyzing presidential debates. *Nordicom Review*, 32(1):31–43, 2011.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1113–1122, 2014.

Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. United Nations General Debate Corpus, 2017. URL https://doi.org/10.7910/DVN/0TJX8Y.

Peter John. The policy agendas project: a review. *Journal of European Public Policy*, 13(7): 975–986, 2006.

Peter John and Will Jennings. Punctuations and turning points in British politics: The policy agenda of the Queen's speech, 1940–2005. *British Journal of Political Science*, 40(3):561–586, 2010.

Kristen Johnson and Dan Goldwasser. Identifying stance by analyzing political discourse on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science (NLP+CSS)*, pages 66–75, 2016.

Kristen Johnson and Dan Goldwasser. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 720–730, 2018.

Kristen Johnson, Di Jin, and Dan Goldwasser. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 741–752, 2017.

Kenneth Joseph, Lisa Friedland, William Hobbs, David Lazer, and Oren Tsur. Constance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1115–1124, 2017.

Willy Jou and Russell J. Dalton. Left-right orientations and voting behavior. *Oxford Research Encyclopedia of Politics*, pages 1–22, 2017.

Amir Karami and Aida Elkouri. Political popularity analysis in social media. In *International Conference on Information (iConference)*, pages 456–465, 2019.

Mladen Karan, Jan Šnajder, Daniela Širinic, and Goran Glavaš. Analysis of policy agendas: Lessons learned from automatic topic classification of Croatian political texts. In *Proceedings of the Workshop on Language Technology for Cultural Heritage, Social Sciencesand Humanities (LaTeCH) at ACL*, pages 12–21, 2016.

Daniel Kauffman, Foaad Khosmood, Toshihiro Kuboi, and Alex Dekhtyar. Learning alignments from legislative discourse. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (DG.O)*, pages 1–2, 2018.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 1–10, 2018.

Angelika Kimmig, Stephen Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pages 1–4, 2012.

John W. Kingdon. *Agendas, Alternatives and Public Policies*. Longman, 1995.

Jan Kleinnijenhuis and Wouter De Nooy. Adjustment of issue positions based on network strategies in an election campaign: A two-mode network autoregression model with cross-nested random effects. *Social Networks*, 35(2):168–177, 2013.

H.D. Klingemann, R.I. Hofferbert, and I. Budge. *Parties, policies, and democracy*. Theoretical lenses on public policy. Westview Press, 1994.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. Domain agnostic real-valued specificity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 6610–6617, 2019.

Ryota Kobayashi and Renaud Lambiotte. TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media (ICWSM)*, pages 191–200, 2016.

Thomas König, Brooke Luetgert, and Tanja Dannwolf. Quantifying European legislative research: using CELEX and PreLex in EU legislative studies. *European Union Politics*, 7(4): 553–574, 2006.

Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman. Party matters: Enhancing legislative embeddings with author attributes for vote prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–515, 2018.

Petia Kostadinova. Party pledges in the news: Which election promises do the media report? *Party Politics*, 23(6):636–645, 2017.

Peter Kraft, Hirsh Jain, and Alexander M Rush. An embedding model for predicting roll-call votes. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2070, 2016.

Gregory J Kramer. *The apathetic country: Are Australians interested in politics and does it matter?* PhD thesis, Queensland University of Technology, 2018.

Nane Kratzke. The #btw17 Twitter dataset–recorded tweets of the federal election campaigns of 2017 for the 19th German Bundestag. *Data*, 2(4):34, 2017.

Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys*, 53(1), February 2020.

Onawa P Lacewell and Annika Werner. Coder training: Key to enhancing reliability and validity. *Mapping Policy Preferences from Texts*, 3:169–194, 2013.

Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 993–1003, 2013.

Virgile Landeiro and Aron Culotta. Robust text classification in the presence of confounding bias. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 186–193, 2016.

Virgile Landeiro and Aron Culotta. Robust text classification under confounding shift. *Journal of Artificial Intelligence Research*, 63:391–419, 2018.

Virgile Landeiro, Tuan Tran, and Aron Culotta. Discovering and controlling for latent confounds in text classification using adversarial domain adaptation. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 298–305, 2019.

Kirsty Laurie, Jason McDonald, et al. A perspective on trends in Australian government spending. *Economic Roundup*, (1):27–49, 2008.

Michael Laver and Ian Budge. The policy basis of government coalitions: A comparative investigation. *British Journal of Political Science*, 23:499–519, 1993.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

Francis LF Lee. Social media, political information cycle, and the evolution of news: The 2017 chief executive election in Hong Kong. *Communication and the Public*, 3(1):62–76, 2018.

Kevin Lerman, Ari Gilder, Mark Dredze, and Fernando Pereira. Reading the markets: forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 473–480, 2008.

Elisabeth Lex, Andreas Juffinger, and Michael Granitzer. Objectivity classification in online media. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, page 293–294. Association for Computing Machinery, 2010. ISBN 9781450300414.

Ralf Lindner and Ulrich Riehm. Broadening participation through e-petitions? An empirical study of petitions to the German parliament. *Policy & Internet*, 3(1):1–23, 2011.

Marco Lippi and Paolo Torroni. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2979–2985, 2016.

L Lipschits. *Verkiezingsprogrammas 1977: verkiezingen voor de Tweede Kamer der Staten-Generaal*. Staatsuitgeverij, Den Haag, 1977.

Dilin Liu and Lei Lei. The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, context & media*, 25:143–152, 2018.

Linqing Liu, Huan Wang, Jimmy Lin, Richard Socher, and Caiming Xiong. Attentive student meets multi-task teacher: Improved knowledge distillation for pretrained models. *arXiv preprint arXiv:1911.03588*, 2019.

James Lo, Sven-Oliver Proksch, and Thomas Gschwend. A common left-right scale for voters and parties in Europe. *Political Analysis*, 22(2):205–223, 2013.

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

Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2):254–277, 2015.

Helen Z Margetts, Peter John, Scott A Hale, and Stéphane Reissfelder. Leadership without leaders? Starters and followers in online collective action. *Political Studies*, 63(2):278–299, 2015.

Theres Matthieß. Retrospective pledge voting: A comparative study of the electoral consequences of government parties' pledge fulfilment. *European Journal of Political Research*, 59:1–23, 2020.

Stefano Menini and Sara Tonelli. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical papers (COLING)*, pages 2461–2470, 2016.

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2938–2944, 2017.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

Burt L Monroe and Philip A Schrodt. Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, 16(4):351–355, 2008.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. Digging in the dirt: Extracting keyphrases from texts with KD. In *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, pages 3–4, 2015.

Lisa Müller. The impact of the mass media on the quality of democracy within a state remains a much overlooked area of study. *LSE European Politics and Policy (EUROPP) Blog*, 2014. URL https://bit.ly/30NE8yJ.

Stefan Müller. Prospective and retrospective rhetoric: A new dimension of party competition and campaign strategies. In *Manifesto Corpus Conference*, 2018.

Stefan Müller. Media coverage of campaign promises throughout the electoral cycle. *Political Communication*, pages 1–23, 2020.

Nona Naderi and Graeme Hirst. Recognizing reputation defence strategies in critical political exchanges. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 527–535, 2017.

Nona Naderi and Graeme Hirst. Using context to identify the language of face-saving. In *Proceedings of the 5th Workshop on Argument Mining (ArgMining)*, pages 111–120, 2018.

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

Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Workshop on 'Making Sense of Microposts': Big things come in small packages (#Microposts)*, pages 93–98, 2011.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media (ICWSM)*, pages 122–129, 2010.

OECD. Trust in government, policy effectiveness and the governance agenda. *Government at a Glance*, pages 19–37, 2013.

Carole Pateman. *Participation and democratic theory*. Cambridge University Press, 1970.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Technical report, The University of Texas at Austin, 2015.

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

François Pétry and Dominic Duval. Electoral promises and single party governments: The role of party ideology and budget balance in pledge fulfillment. *Canadian Journal of Political Science/Revue canadienne de science politique*, 51(4):907–927, 2018.

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

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference (WWW)*, pages 155–158, 2018a.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22–32, 2018b.

Julia Proskurnia, Karl Aberer, and Philippe Cudré-Mauroux. Please sign to save...: How online environmental petitions succeed. In *The Workshops of the Tenth International AAAI Conference on Web and Social Media Social Web for Environmental and Ecological Monitoring*, pages 195–198, 2016.

Julia Proskurnia, Przemyslaw Grabowicz, Ryota Kobayashi, Carlos Castillo, Philippe Cudré-Mauroux, and Karl Aberer. Predicting the success of online petitions leveraging multidimensional time-series. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pages 755–764, 2017.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. An automated method of topic-coding legislative speech over time with application to the 105th-108th US senate. In *Midwest Political Science Association Meeting*, 2006.

Ludovic Rheault. Expressions of anxiety in political texts. In *Proceedings of the First Workshop on NLP and Computational Social Science (NLP+CSS)*, pages 92–101, 2016.

Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2): 107–136, 2006.

Robert E Riggs. The United Nations as an influence on United States policy. *International Studies Quarterly*, 11(1):91–109, 1967.

Martin Ringsquandl and Dusan Petkovic. Analyzing political sentiment on Twitter. In *Analyzing Microtext: 2013 AAAI Spring Symposium Series*, pages 40–47, 2013.

David Bruce Robertson. *A theory of party competition*. John Wiley & Sons, 1976.

Terry J Royed. 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, 1996.

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

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Zaher Salah, Frans Coenen, and Davide Grossi. Extracting debate graphs from parliamentary transcripts: A study directed at UK House of Commons debates. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (ICAIL)*, pages 121–130, 2013.

Prathusha Kameswara Sarma, Yingyu Liang, and Bill Sethares. Domain adapted word embeddings for improved sentiment classification. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, pages 51–59, 2018.

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

Philippe C. Schmitter and Alexander H. Trechsel. Green paper : The future of democracy in Europe. Trends, analyses and reforms. https://www.coe.int/t/dgap/democracy/activities/key-texts/02_Green_Paper/GreenPaper_bookmarked_en.asp, 2004. Accessed: 2020-01-20.

Mirco Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. The UN Security Council debates 1995-2017. *arXiv preprint arXiv:1906.10969*, 2019.

Gijs Schumacher, Catherine E De Vries, and Barbara Vis. Why do parties change position? party organization and environmental incentives. *The Journal of Politics*, 75(2):464–477, 2013.

J. R. Searle. *A Taxonomy of Illocutionary Acts*. Linguistic Agency University of Trier, 1976.

Matthew A. Shapiro, Libby Hemphill, and Jahna Otterbacher. Updates to congressional speech acts on Twitter. In *Proceedings of the American Political Science Association Annual Meeting*, 2018.

Kiran Sharma, Gunjan Sehgal, Bindu Gupta, Geetika Sharma, Arnab Chatterjee, Anirban Chakraborti, and Gautam Shroff. A complex network analysis of ethnic conflicts and human rights violations. *Scientific reports*, 7(1):1–7, 2017.

Aili Shen, Bahar Salehi, Jianzhong Qi, and Timothy Baldwin. Feature-guided neural model training for supervised document representation learning. In *Proceedings of the The 17th*

Annual Workshop of the Australasian Language Technology Association (ALTA), pages 47–51, 2019.

Jeremy Shiffman. Agenda setting in public health policy. In *International encyclopedia of public health*, pages 16–21. Elsevier Inc., 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniela Širinić. Political attention of Croatian governments 1990–2015. In *Policy-Making at the European Periphery*, pages 65–82. Springer, 2019.

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

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2895–2905, 2019.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Systems Research and Behavioral Science*, 7(4):484–498, 1962.

Eva Strangert. Prosody in public speech: analyses of a news announcement and a political interview. In *Proceeding of the Ninth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 3401–3404, 2005.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 3319–3328, 2017.

Chenhao Tan, Hao Peng, and Noah A Smith. "You are no Jack Kennedy" on media selection of highlights from presidential debates. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pages 945–954, 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1195–1204, 2017.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 327–335, 2006.

Robert Thomson. 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, 2001.

Robert Thomson, Terry Royed, and Elin Naurin. The Program-to-Policy Linkage: A Comparative Study of Election Pledges and Government Policies in the United States, the United Kingdom, the Netherlands and Ireland. In *Proceedings of the American Political Science Association Annual Meeting*, 2010.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 809–819, 2018.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth international AAAI conference on weblogs and social media (ICWSM)*, pages 178–185, 2010.

Andrew Turpin. An attempt to measure the quality of questions in question time of the Australian federal parliament. In *Proceedings of the Seventeenth Australasian Document Computing Symposium (ADCS)*, pages 96–103, 2012.

Peter Van Aelst and Rens Vliegenthart. Studying the Tango: An analysis of parliamentary questions and press coverage in the Netherlands. *Journalism Studies*, 15(4):392–410, 2014.

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

Bertie Vidgen and Taha Yasseri. What, when and where of petitions submitted to the UK government during a time of chaos. *Policy Sciences*, pages 1–23, 2020.

David Vilares and Yulan He. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1573–1582, 2017.

Petra Vodová. Who makes a compromise? Adopting pledges in Czech coalition agreements. *European Review*, pages 1–16, 2020.

Andrea Volkens, Onawa Lacewell, and Henrike Schultzeand Annika Werner. Pola Lehmann, Sven Regel. The manifesto data collection. Manifesto Project (MRG/CMP/MARPOR). In *Wissenschaftszentrum Berlin fur Sozialforschung (WZB)*, 2011.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. AllenNLP Interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, 2019a.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5310–5318, 2019b.

Annemarie S Walter. Choosing the enemy: Attack behaviour in a multiparty system. *Party Politics*, 20(3):311–323, 2014.

Zhongyu Wei, Yulan He, Wei Gao, Binyang Li, Lanjun Zhou, and Kam-fai Wong. Mainstream media behavior analysis on Twitter: a case study on UK general election. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT)*, pages 174–178, 2013.

Rebecca Weitz-Shapiro and Matthew S Winters. Political participation and quality of life. Technical report, Working Paper, 2008.

Annika Werner, Onawa Lacewell, and Andrea Volkens. Manifesto coding instructions (4th fully revised edition). *URI:http://goo.gl/g512Q*, 2011.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 347–354, 2005.

Woodrow Wilson. *Responsible Party Government, 1885*, pages 163–167. Palgrave Macmillan US, New York, 2002.

Leah Cathryn Windsor, James Grayson Cupit, and Alistair James Windsor. Automated content analysis across six languages. *PLOS ONE*, 14(11), 2019.

John T Woolley and Gerhard Peters. The American Presidency Project. Santa Barbara, CA. https://www.presidency.ucsb.edu/, 2011.

Scott Wright. E-petitions. In *Handbook of digital politics*. Edward Elgar Publishing, 2015.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2019.

Hao Yan, Allen Lavoie, and Sanmay Das. The perils of classifying political orientation from text. *Linked Democracy: Artificial Intelligence for Democratic Innovation*, 858(8), 2017.

Yiming Yang, Jaime G Carbonell, Ralf D Brown, Thomas Pierce, Brian T Archibald, and Xin Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

Tae Yano, Noah A Smith, and John D Wilkerson. Textual predictors of bill survival in congressional committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 793–802, 2012.

Taha Yasseri, Scott A Hale, and Helen Z Margetts. Rapid rise and decay in petition signing. *EPJ Data Science*, 6(1):20, 2017.

Elyse Yates and Sherri Greenberg. Social media, legislation and bringing the public inside. In *Proceedings of the 15th Annual International Conference on Digital Government Research (DG.O)*, pages 314–315, 2014.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*, 2018.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. Asking too much? the rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1558–1572, 2017.

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 417–424, 2011.

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