From Political Manifestos to Social Networks: the automation of Political Discourse Analysis using Contextual Information

PhD dissertation by Aritz Bilbao Jayo

within the doctoral program Engineering for the Information Society and Sustainable Development

Candidate	Advisor
Aritz Bilbao Jayo	Dr. Aitor Almeida

Lehenengo Aitari, hau ikustea asko gustatuko zitzaiolako
Aitxitxe ta Amumari, bidaiaren hasieran joan zirenak
Nire Ama, Arreba eta Arianeri, beti egoteagatik
Eta azkenik nire Iloba Uxueri, bidean agertzeagatik.

Abstract

Due to the rise of the social networks, political parties and politicians have found new ways of establishing their position on an issue apart from traditional political manifestos. From this phenomenon, a new research area has emerged, the automation of political discourse analysis on Social Networks. To do so, this PhD dissertation has taken advantage of a widely used content analysis methodology for political manifestos, The Manifesto Project. With annotated manifestos since 2001, this methodology uses a codification which allows the analysis of political parties policy preferences regarding 56 topics, providing the scientific community with parties' policy positions derived from the content analysis.

Therefore, this PhD dissertation focuses on two main tasks: firstly, to automate the annotation process of political manifestos, in order to facilitate that same process to political scientists and secondly, to use this model as a basis to perform a political discourse analysis on Twitter using the previously mentioned Manifesto Project's methodology. To do so, we have taken advantage of two types of contextual information available in the two circumstances of the application of this research work: manifestos and Twitter. The first contextual data is what has been said previously, in the case of election manifestos the previous phrase or statement, and on twitter the preceding tweet. The second contextual information is which political party is the sender of the statement.

Regarding the use of contextual information in order to improve manifestos automated classification, we have improved state of the

art results in 4 out of 7 languages. With regard to Tweets' classification, we can affirm that annotated manifestos can be used as complementary data for this task, being the fine-tuned model with annotated tweets the best performing one. Moreover, contextual information does also improve the performance of the models when tweets are classified. Using this approach, we have analysed the 2016 United States presidential elections on Twitter.

Resumen

Debido al auge de las redes sociales, los políticos y sus respectivos partidos han encontrado nuevas formas de dar a conocer su posición sobre cualquier tema, en otros lugares que no son los programas electorales. De este fenómeno, ha emergido una nueva área de investigación, la automatización del análisis del discurso político en redes sociales. Para ello, en esta tesis doctoral se hace uso de una metodología diseñada por el Manifesto Project para el análisis de contenido en programas electorales. Con programas manualmente anotados desde 2001, este proyecto propone una codificación que permite identificar las preferencias políticas de los partidos políticos con respecto a 56 categorías diferentes, proveyendo a la comunidad científica las posiciones de estos partidos tras aplicar técnicas de análisis de contenidos.

Por tanto, esta tesis doctoral se centra es dos tareas: primero, automatizar el proceso de anotación de programas electorales, para así facilitar este mismo trabajo a los politólogos y segundo, usar este modelo como base para analizar el discurso político en redes sociales usando la metodología del Manifesto Project. Para ello, hemos usado el contexto disponible en las dos circunstancias donde se va a aplicar esta investigación: los programas electorales y Twitter.

El primer tipo de contexto usado es qué se ha dicho anteriormente, en el caso de los programas la frase o afirmación previa y en Twitter el tuit anterior. El segundo es el partido político que ha realizado la afirmación.

Respecto al uso de información contextual para mejorar la clasificación automática de programas electorales, hemos mejorados los resultados del estado del arte en 4 de 7 idiomas. Con respecto a la clasificación de tuits, podemos afirmar que los programas electorales anotados puede ser usados como información complementaria para esta tarea, siendo el modelo reentrenado con tuits anotados el que mejor resultados obtiene. Además, la información contextual también mejora el rendimiento del modelo. Por último, usando este mismo enfoque, hemos analizado las elecciones presidenciales estadounidenses de 2016 en Twitter.

Acknowledgements

Esta Tesis Doctoral se ha realizado con la ayuda y el apoyo de mucha gente que me ha estado aguantando durante todos estos años. Baina lehenik eta behin, nire aitari eskerrak ematea gustatuko litzaidake, nigan betin sinetsi zuelako eta niretzako garrantzi gabeko lorpenak zirenak, berarentzat munduko lorpen garrantzitsuenak zirelako. Eta ziur nago, nahiz eta nik, tesi hau ez delako hainbesterako esan, berarentzako munduko tesirik onena izango zelako.

Nire Ama, Arreba eta Arianeri 3 urte hautean nire umore txarrak eta gorabeherak jasan dituztelako eta laguntzeko prest egon direlako beti. Eskerrik asko.

También me gustaría agradecer a mi director Aitor Almeida, que ha aguantando con mucha paciencia todas mis dudas, inseguridades, bloqueos, etc. Sin obviar, la ayuda que me ha dado a la hora de tener ideas para esta tesis y por lo que ha luchado para que siga en Deustotech.

A mis amigos de siempre Jon Lorente, Peio Iñurrigarro y Joseba Rojo (Futxin) que me han entendido durante estos años. Así como a todo el grupo de MORELab: Mikel Emaldi, Unai Aguilera y Anne Miren por ayudarme a relativizar las cosas, Adrian (mi compañero de tesis) por ser un ejemplo de actitud y motivación, y a los minions Zulaika y Rubén por el buen rollo que trajeron.

Also, I would like to mention my research stay in the University of Warwick which helped me in the definition of this dissertation, understanding what is science and what is not.

Por último, al grupo de trastornados de Vesania, por todas las risas que nos hemos echado durante el último año.

Thank you
Aritz Bilbao Jayo

Table of Contents

List of Figures

List of Tables

Acronyms

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

BOE Boletin Oficial del Estado

CMP Comparative Manifestos Project

CNN Convolutional Neural Networks

GLES German Longitudinal Election Study

LM Language Model

LSTM Long short-term memory

MLM Masked Language Modelling

MRG Manifesto Research Group

NLP Natural Language Processing

RMP Regional Manifestos Project

Ekin ta aurrera.

Bittor Joseba Bilbao

CHAPTER

1

Introduction

Over the last decades political scientists have been analysing election manifestos in order to perform a discourse analysis of political parties using their respective manifestos. This method has allowed them to perform several studies using content analysis techniques. These researches usually involve studies such as a temporal or longitudinal analysis of how parties' political discourse has evolved over the years, taking only into account their election manifestos (Benoit, 2009), the comparison between different type of manifestos (national and European level manifestos (Wüst and Volkens, 2003)), analysing how much parties emphasize certain topics and which are their positions in some specific topics depending on the elections context(Alonso et al., 2017) or to estimate policy positions for political parties on left-right scales using measures such as RILE scale (Budge, 2013) or other alternatives (Lowe et al., 2011).

To do so, political scientists have been using methodologies such as the one proposed by the Comparative Manifestos Project (CMP)(Budge, 2001). This methodology consists in annotating political manifestos' sentences with a category (among a total of 56 categories), indicating the main idea behind the statement. For instance, *"We will legislate to require all major parties to have their manifesto commitments independently audited by the Office for*

Budget Responsibility" has the category *"Economic Planning: Positive"* or *"Too much power is unaccountable, concentrated in the market and the state, at the expense of individuals and their communities"* with the category *"Democracy"*. It may also happen that a sentence in a manifestos contains more than one idea, in that case, the annotator has to unitise or divide the sentence in various quasi-sentences, one per idea. This unitisying process is explained in depth in Section 3.1.

Nowadays, the category scheme for manifestos annotation consists in 56 categories grouped into seven major policy areas(Volkens et al., 2019) (see table 3.3 in Section 3.1): *External Relations*, *Freedom and Democracy*, *Political System*, *Economy*, *Welfare and Quality of Life* and *Social Groups*.

However, election manifestos have stopped being the unique reference venue where political parties and politicians express their ideas and promises. Instead, Social Networks have started to complement or replace traditional political manifestos, offering a direct means of communication where political ideas and promises are written down in a similar way as political manifestos have traditionally done. In this new context, politicians from all over the world spread everyday a considerable amount of statements using social networks, in contrast to political manifestos which are only written by political parties when elections are held. Therefore, the amount of information that political scientists have to analyse has increased significantly in the last few years. This new phenomenon has created a new research opportunity: to automate the political text annotation process used by political scientists to apply it to online social networks. Moreover, since this is not an easy task even for trained political scientists as (Mikhaylov et al., 2012) demonstrated, due to the high number of available categories for annotation, the automation of this task could benefit political scientists, reducing the amount of time and effort needed for this task.

In this PhD dissertation we aim to address this problem, first to help political science researchers in the annotation process and second, to automatically analyse politician's discourse on social networks (in this case, Twitter). For this purpose, we have tackled the introduced research opportunity as a two steps problem: to help in the automation of manifestos annotation process

and then, use the acquired knowledge and apply it in social networks. Therefore, we have taken advantage of two types of contextual information available in the two circumstances of application of this PhD dissertation. The first contextual data is what has been said previously, in the case of election manifestos the previous phrase or statement, and on twitter the preceding tweet. The second contextual information is which political party has asserted a statement. However, how the political party should be represented as input feature introduces an additional challenge. In this work, we have analysed 4 different party representation methods using diverse information about political parties: using the RILE scale which gives a score in a right-left axis, parties' political orientation and assuming that each party is unique. The best results has been achieved with a disentangled representation(Bengio et al., 2013) based on parties' political orientation, whereas the worst results has been obtained with the left-right scale. Moreover, political orientation based representation allows the addition of new parties to the model without the need of retraining the model again with the new parties.

Regarding the use of contextual information in order to improve manifestos automated' classification, we have improved state of the art results in 4 out of 7 languages with this approach. Then, we have applied the same architectures to tweets in order to analyse its performance classifying tweets in three different scenarios: using a model exclusively trained with election manifestos, with only few annotated tweets or fine-tuning with annotated tweets a model previously trained with manifestos. Among our findings we can affirm that annotated manifestos can be used as complementary data for tweets classification, being the fine-tuned model the best performing one and that contextual information does also improve the performance of the models when tweets are classified.

Finally, we have used for the first time the CMP's codification to analyse politicians' political discourse in Twitter. The 2016 United States presidential elections has been analysed using the 56 categories from the CMP, reaching interesting conclusions and achieving similar results other political scientists have obtained regarding partisan rhetoric in Twitter, which proves these results are aligned with those obtained by other political scientists.

The remainder of the chapter is structured as follows: Section 1.1 introduces the context and motivation of this dissertation. Then, Section 1.2 formulates the hypothesis, objectives and scope of this research. After, Section 1.3 explains the followed methodology. Section 1.4 summarises the main scientific and technical contributions. Finally, Section 1.5 describes the outline of this PhD dissertation.

1.1 Context and motivation

The rise of social networks have offered both politicians and citizens new ways of interacting directly with each other without the direct mediation of traditional media. This phenomenon has allowed citizens to become participants in the construction of the political agenda, forcing political parties to use more direct means of communication than the mainstream press and media. The most representative element of this paradigm is Twitter. Created in 2006, this social network has become one of the most important forms of communication between politicians and their electorate, reaching the point where some politicians bypass traditional media and exclusively release statements on social media. Furthermore, as all the members of the social network are treated as equals, any citizen can send a message to the politician, leading sometimes to a discussion between the politician and the citizens or between citizens.

Therefore, social networks contain valuable data regarding citizens' concerns or the politicians' current talking points. However, since thousands of messages are created every hour, it is not feasible to manually analyse them. Thus, in order to analyse the political discourse in real time, the data analysing process has to be automated. For that purpose, we want to adopt a multidisciplinary approach, to build a text categorization classifier using political manifestos which has been manually annotated by domain experts, combining the political science knowledge from the social scientists involved in the annotation of the political manifestos with natural language processing, in order to be able to process large quantities of data and study how the political discourse evolves online, in this case, Twitter.

Originated from this new opportunity, the research community started designing innovative tools or approaches to automatically analyse anything that occurs on social networks, being political opinion mining one of the main research fields. As it is extensively explained in Section 2, several approaches have been designed over the last decade with a wide range of objectives, from the various attempts to predict political elections' results, to the political polarity analysis (mainly left-right axis) or plain sentiment analysis of some specific topics or politicians. However, to the best of our knowledge, none of the reviewed approaches have used in Social Networks a widely adopted methodology by political scientists which already has been proven to be successful when it comes to analyse political discourse on several topics and axes, and therefore, overcoming the traditional analysis on the left-right axis.

Moreover, the approach introduced in this dissertation relies on the advantages that a multidisciplinary perspective could contribute in this field, combining manually annotated political manifestos by political scientists with advanced natural language processing techniques and the contextual information that we believe annotators used when it comes to codifying election manifestos: what it has been said previously and who has said it.

Thus, we believe that applying in Social Networks an already validated methodology for manifestos analysis, where politicians have found a new place to spread their ideas, is a robust starting point in order to automatically perform more complicated or profound analyses than the traditional studies based on polarization (pro-against, left-right, republicans-democrats) or content analysis works focused on some particular topics. Moreover, apart from the fact that CMP's methodology has been extensively used since early 2000, this has allowed the creation of a large dataset of annotated manifestos that we consider essential in order to apply CMP's methodology in a new area such as Social Networks where does not exist any dataset with annotated tweets or Facebook posts with this codification. Thus, the initial work (to annotate political text with the CMP's codification) is already done and we can take advantage of it, in order to build our tool for political discourse analysis.

1.2 Hypothesis, Objectives and Scope

Based on the current state of automated use of political manifestos for their automated codification or social network political discourse analysis, the hypothesis of this dissertation is:

> **Hypothesis 1** *Using contextual information it is possible to improve the automated election manifestos annotation process and perform a political discourse analysis in on-line social networks using manifestos' annotation scheme and the same contextual data previously used.*

To be able to validate this hypothesis the general goal of this research project is:

> **Goal 1** *To design and implement a political discourse classifier that uses annotated political manifestos, a very reduced amount of annotated political tweets and the context of each of those tweets to analyse on-line political discourse.*

This general goal can be achieved by addressing the following and therefore, more specific and measurable objectives.

1. To study the current start of the art on political discourse analysis in Social Networks and the automated used of annotated political manifestos.

2. To design and implement a deep learning supervised classification model for text categorization optimized for the problem and able to have different inputs than raw text.

3. To identify an appropriate evaluation methodology for the automated manifestos annotation task with its corresponding metrics and perform a quantitative analysis of the results.

4. To analyse how the added contextual data affects supervised classifier's performance when classifying election manifestos.

5. To identify an appropriate evaluation methodology for the on-line political discourse analysis task with its corresponding metrics and perform a quantitative analysis of the results.

6. To analyse if annotated political manifestos could be used as complementary data to the annotated tweets in order to improve the performance of the political discourse classifier.

7. To analyse how does the designed approach analyse the on-line political discourse using contextual information.

The resulting political discourse analyser system should also fulfil the following requirements:

1. Language independence: the developed political discourse classifier should be able to be modelled in any language as long as there are annotated political manifestos in that language.

2. The designed architecture should be able to accept new types of contextual information at any moment.

The work presented in this dissertation does not deal with the following conditions:

1. We assume that the text has already been unitised when it has to be classified. The unitising process is out of the scope of this dissertation.

2. Even though the Manifestos Codification schema has been extended by various projects, during this dissertation Manifestos Project's codification schema has been used.

1.3 Methodology

The following strategy (see Figure 1.1) has been followed in order to accomplish the hypothesis and goals presented in Section 1.2.

1. **Exploratory phase**: explore the literature related to the research field in order to build a solid theoretical framework on which support the rest of the work. Even though the task of revising the literature is presented as an isolated task, it is clear that it is an incremental and continuous process and consequently it will be done throughout the entire project.

2. **Definition of the validation scenario and conceptual test**: after the first phase of the study of the state of art, and knowing the limitations and advantages of the research proposal, a first version of the scenario in which the project will be developed will be defined. Even though it is expected to evolve during the investigation, having defined a strong initial framework will help steer the research towards the analysis that form the basis of it.

3. **Specification and design of the solution**: at this stage, the necessary requirements to solve the initial starting point will be specified, and the solution that best results can achieve on it will be designed. To do so, it is essential to continue gathering information on the state of the art, so that any innovation in the field will be considered in the design if it were relevant.

4. **Design of the test and evaluation**: after the design, the testing and evaluation system under which the solution will be assessed will be determined, verifying its validity and applicability in real environments.

5. **Development of a functional prototype, dissemination and writing the PhD dissertation**: the final task of the research will focus all efforts on the development of a demonstrator or functional prototype to justify each of the above tasks. In addition, the results obtained are expected to be innovative and of great significance for both academic and social fields. Ultimately, this thesis will be finished and refined for its later submission and defence.

1.4 Contributions

The following scientific contributions can be found in this dissertation:

- When it comes to the automated codification of political manifestos, it has been statistically certified that, adding the previous phrase improves the performance of the classifier.
- Also, it has been statistically certified that, using as contextual information the political party to which the phrase belongs, improves the performance of the classifier, particularly when classifying subdomains.
- A novel method for political parties representation has been designed using disentangled representation and parties' political orientation.
- It has been proven how annotated political manifestos and annotated political tweets are complementary information when it comes to training the political discourse classifier.
- It also has been proven that using the previous tweet and political party as additional contextual data achieves the best results classifying annotated tweets.
- A novel approach for automatically classifying political tweets using a categorisation scheme widely used by political scientist.
- A dataset of 5,000 tweets annotated with the CPM coding schema has been created.

The following technical contributions can be found in this dissertation:

- Word2Vec embedding models for the Spanish language from text recovered from news, Wikipedia, the Spanish BOE, web crawling and open literary sources with a total of 3.257.329.900 words and 18.852.481.207 characters. (Almeida and Bilbao, 2018)

1.5 Thesis outline

This PhD dissertation is structured in 6 chapters. The current section, Chapter 1, introduces the dissertation explaining its context, motivation, objectives, methodology and contributions.

Chapter 2 presents an analysis of the state of the art relevant for this dissertation.

Chapter 3 describes the theoretical foundations of the used political discourse analysis methodology, on which the research work conducted during this dissertation has been constructed.

Chapter 4 presents the used contextual data. The reason behind their addition, how they have been added, etc. Also, the used machine learning models and how they have been modified for this task is explained.

Chapter 5 describes the used evaluation methodology and the achieved results during this dissertation.

Chapter 6 summarises the main findings and contributions of this PhD dissertation and proposes future work.

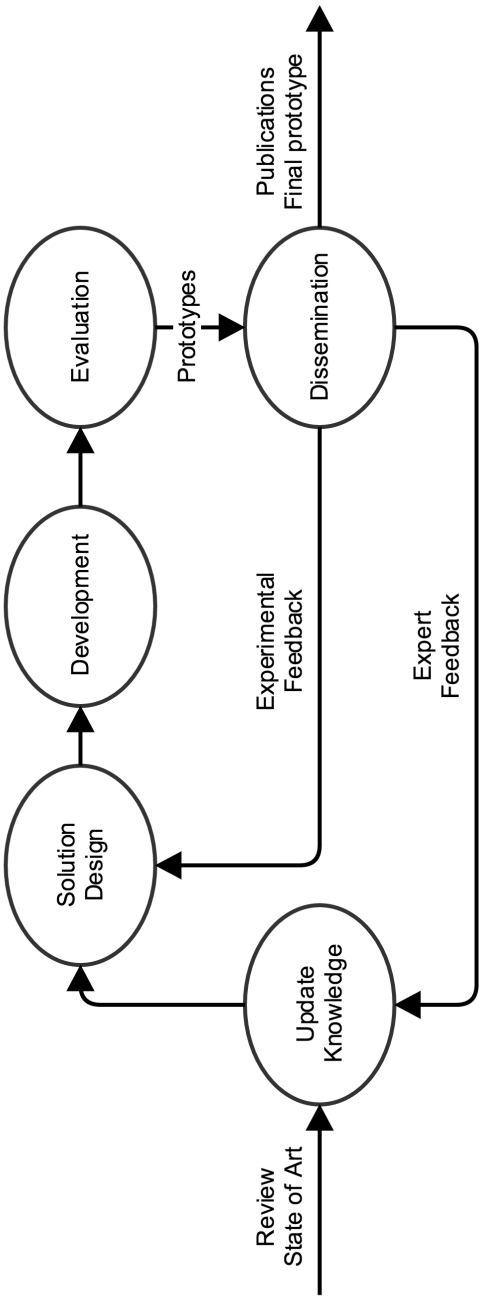

Figure 1.1: Research Methodology

Jokatzeko ordua heldu da. Aurrenen egin behar dudana neure buruaz fidatzea da eta neure larritasuna menperatu.

Son Goku

CHAPTER

2

Related work

SINCE its inception, Twitter has been seen by researchers of several fields as a new source of information with which they can conduct their researches. For instance, political scientists have identified Twitter as a platform where they can analyse what a subset of the population says without performing expensive surveys, study how politicians prioritise some topics over others or which ideas politicians want to send to their followers. This phenomenon has offered to political scientists, on the one hand, and computer scientists on the other, a new research opportunity. In the first case, political science researchers have focused their work in manual approaches where each message or statement is manually analysed by a human, to later drawn some conclusions having as basis those manually annotated messages. On the contrary, computer scientists have taken a more automated approach where different aspects of the tweets are automatically analysed to later drawn conclusions from them. Either way, both approaches have led to a large number of research publications. In this PhD dissertation we have aimed to combine both worlds and therefore, we have divided state of the art chapter in two parts: first, we have reviewed the most relevant works manually analysing the political discourse in social networks and second, those research works using automated approaches. In Chapter 3,

all the related work regarding the automated used of political manifestos is explained.

2.1 Manual approaches

When it comes to manual approaches, we refer to those political analyses made with social media data (which have been probably gathered automatically using APIs or crawlers), but each of the posted message or tweet has been analysed manually, without the intervention of any supervised or semi-supervised tool.

Even though this may be seen out of the scope of this PhD dissertation, we believe this an interesting study in order to show the potential that automation or at least, the semi-automation of these processes would have.

(Ramos-Serrano et al., 2018) analysed the twitter activity of Spanish political parties during the 2014 European campaign. They manually analysed questions such as with whom are Spanish parties interacting, which topics are they tweeting about or what was the function of politicians tweets. Authors found that Spanish conservative parties retweeted less messages than the rest of the parties, while new parties retweeted messages the most. When it comes to replying to others users, the two traditional and majoritarian parties, PSOE and PP, were the parties who less replied. Regarding the type of user with whom parties interact, Podemos was the party that dialogued the most with citizens, whereas most of the replies were directed to politicians and in a very low percentage to journalists. With regard to the topics of the tweets, *"Campaign and Party Affairs"* was the main topic. Then, topics such as *"Europe"*, *"Corruption"* (mainly by minor parties) and *"Nationalism"* (by centre-right parties) were the most treated during campaign. In total, authors analyses 21 different topics.

(López-García, 2016) studied the 2015 Spanish general elections campaign in Twitter. In particular, the research work focuses on analysing main political parties candidates from three perspectives: a quantitative analysis of the tweets; focusing on the number of responses and retweets, a content analysis in order to study the agenda of each candidate and finally a qualitative

analysis where the communication preferences of each candidate are studied. The content analysis consisted in classifying politicians tweets in 4 different categories: political, policy, campaign and personal, far from the CMP's 56 categories designed for political content analysis.

(Casero-Ripollés et al., 2017) analysed the messages sent by the political party Podemos (candidates and official Twitter accounts). The authors performed a quantitative analysis focusing on the issues and functions of the messages sent by the party. On one hand, in order to study the functions, the authors created an ad hoc taxonomy of 13 categories: agenda and organization of political actions, electoral program, management of political achievements, criticizing opponents, media agenda, interaction and dialogue with users, participation and evaluation, values and ideology, personal life, entertainment, humour, manners and protocol, and others. On the other hand, another taxonomy of 18 elements was created in order to study the topics of the tweets. Among the most relevant categories were: economy, social policy, science and technology, state territorial model, or relationship with the media.

(Russell, 2018) studied the U.S. Senators party polarization identifying those messages with a partisan rhetoric. To do so, Russell catalogued U.S. senators Twitter activity during the first 6 months of the 113th (Democratic majority) and 114th (Republican majority) congresses reaching interesting outcomes. (Russell, 2018) analysed two congresses with different majorities expecting changes in political parties' rhetoric. As it is stated in (Russell, 2018), when this PhD dissertation is being written, the political situation in the United States is highly party-polarized. Having this a fact, (Russell, 2018) categorised tweets sent by Democrat and Republican senators in the previously mentioned period in a partisan, non-partisan classification. To clarify, partisan rhetoric could be defined as those statements praising their own political parties or criticising the opponent's parties. In 2013, with the democrats as majority, 17.3% of Republicans tweets contained partisan rhetoric, unlike Democrats, where 4.5% included this rhetoric. In 2015, even though the majority shifted towards Republicans, they maintained as the party with most partisan rhetoric, 11.75%, in contrast of the 5.43% of Democrats messages. Moreover, the manuscript analysed if those partisan tweets were positive or

negative, concluding that two thirds of Republicans' partisan tweets included negative rhetoric, this percentage decreased to 50% with Democrats. All these partisan rhetoric is related to the *Political Authority* category in CMP which is used in order to analyse the political discourse in Section 5.5.

To sum up, all the reviewed manual approaches have used different categorisation schemes for several purposes with diverse number of topics, being most of those schemes created ad hoc for the research: 21 in (Ramos-Serrano et al., 2018) , 4 in (López-García, 2016) or 13 (Casero-Ripollés et al., 2017). All these schemes are far from the 56 categories of the CMP, which have been already used for manifestos analysis and offer a more in-depth analysis due its low level granularity.

2.2 **Automated approaches**

Automated approaches has been used for several tasks related to political analysis in Social Networks.

One of the main research areas has consisted in the measurement of the predictive power of social networks such as Twitter: predicting election results or comparing opinions from Twitter users on some specific topics, in contrast with real polls regarding the same topic. For instance, (Tumasjan et al., 2010) claimed that the mere number of messages mentioning a party reflects the election result. Tumasjan et al. analysed the 2009th German federal election using more than 100,000 tweets published during election campaign. The authors selected the tweets mentioning a particular party and compared the distribution of tweets per political party with their results on the elections. Finally concluding that their approach's predictive power is close to the classical election polls. Moreover, (O'Connor et al., 2010) measured the potential of Twitter messages as a substitute of traditional polling. 0'Connor et al. gathered tweets about 3 different topics: consumer confidence, presidential approval and elections. Then, authors gave sentiment scores to each of the messages by counting positive and negative words using a subjectivity lexicon and assigning a sentiment score to each day. 0'Connor et al. concluded

that a simple sentiment analysis on top of Twitter data produces similar results to consumer confidence and presidential job approval polls: suggesting that more advanced natural language processing techniques could improve its opinion estimation.

However, criticisms regarding the predictive power of Twitter to forecast elections or use it as a substitute of polls have emerged. (Gayo Avello et al., 2011) replicated Tumasjan et al.'s and O'Connor's approaches utilising a set of tweets about the 2010 United States House of Representatives elections. Compared to Tumasjan et al., Gayo et al. obtained a mean average error of 17.1% compared to election's real results. A greater Mean Average Error than the 1.65% MAE obtained by Tumasjan et al. on 2009th German federal election. Concerning O'Connor's lexicon-based sentiment analysis approach, Gayo et al. concluded that whenever this approach is applied to political conversation, its performance is poor

The analysis of political polarization in social networks has also been an important research field in political activity in Social Networks. There are several studies which have been able to detect the polarity or political orientation regarding an event, idea or political party of Twitter's users. To this effect, one of the main approaches to analyse the polarity in Twitter is to construct the graph representation of the social network and apply some principles of network theory. On one hand, (Conover et al., 2011) used a combination of community detection algorithms and manually annotated data to analyse the polarity of two networks constructed after gathering more than 250,000 tweets about 2010 U.S congressional midterm elections. The first network represented the retweets and the second one the mentions between different users. Conover et al. concluded that users tend to retweet tweets of users they agree with. Therefore, communities are evident in the retweet network. However, in the mentions network there were more interactions between people with different political ideas, suggesting the existence of discussions between different polarities. In consequence, communities of users with same polarity are not as clear as they are in the retweet network. On the other hand, (Finn et al., 2014) presented a new approach for the measurement of the polarity: 'a co-retweeted network that represents how many times two users have both

been retweeted by other users'. They tested their approach with the most retweeted 3,000 tweets within their dataset. They claimed that by using their co-retweeted network were able to measure the polarity of the most important accounts participating in the discussion and the polarity of the analysed event. Other works have detected the polarity of raw text using natural language processing techniques. (Iyyer et al., 2014) designed a recursive neural network in order to identify the political polarity of a sentence. Iyyer et al. used two datasets to evaluate their model: an existing one and another one annotated by them by means of crowdsourcing. Authors were able to identify most conservative or liberal n-grams and detect bias more accurately. Similar works have been conducted on Twitter, such as (Rao and Spasojevic, 2016), in which Rao et al. used word embeddings and LSTM recurrent neural networks in order to classify twitter messages as democratic or republicans. Authors established the ground truth using Twitter Lists, where users are categorised in different groups (democratic or republican) by other users.

Other researches have detected the polarity of raw text using natural language processing techniques. (Iyyer et al., 2014) designed a recursive neural network in order to identify the political polarity of a sentence. Iyyer et al. used two datasets to evaluate their model: an existing one and another one annotated by them by means of crowdsourcing. Authors were able to identify most conservative or liberal n-grams and detect bias more accurately. Similar works have been conducted on Twitter, such as (Rao and Spasojevic, 2016), in which Rao et al. used word embeddings and Long Short-Term Memory (LSTM) recurrent neural networks in order to classify twitter messages as democratic or republicans. Author established the ground truth using Twitter Lists, where users are categorised in different groups (democratic or republican) by other users.

However, all the analysed approaches rely on data created in its entirety in Twitter. Unfortunately, this data gathered from Twitter could have been manipulated by third party actors or institutions. As (Ratkiewicz et al., 2011) introduce in their study about astroturfing in political campaigns on Twitter, there are individuals whose objective is to launch controlled campaigns in favour or against a precise political organization, candidate or idea using

centrally-controlled accounts. So as to detect those campaigns, Ratkiewicz et al. have designed a machine learning framework combining as they say: 'topological, content-based and crowdsourced features of information diffusion networks'. Other researchers have worked detecting rumours in social networks which may introduce new topics of conversation to the network or influence user's opinion about some subjects. For instance, Zubiaga et al. (Zubiaga et al., 2016b) have designed a methodology for collecting, identifying and annotating rumours in Twitter allowing them to analyse how rumours modify the conversation and how they evolve over time. From this work, Zubiaga et al. generated a dataset which later have used in order to classify tweets related to a rumour in four categories(Zubiaga et al., 2016a): supporting, denying, questioning or commenting.

Therefore, in this PhD dissertation we have focused our work on the most reliable political data Twitter contains: messages sent by politicians and political parties. Thus, from now on, the state of the art analysis will be centred on research works using this type of reliable data.

The first example of this kind of political analysis on Twitter using reliable data is (Stier et al., 2018).Authors analysed the 2013 German federal election campaign in Twitter and Facebook, studying how aligned are the topics discussed by politicians compared to the most important topics for the electorate according to a survey, and how their communication strategy vary depending on the Social Network where ideas are spread. To do so, the authors classified the tweets on topics using a human-interpretable Bayesian language model. The topics were defined by known survey classes and additional social-media-specific topics. They used the German Longitudinal Election Study Survey that collected the opinion of 7,882 people before and after elections. In particular, Stier et al. coded the open-ended responses with GLES(Schmitt-Beck et al., 2009) categorisation schema which consists in three high level dimensions, politics, polity and policy, ending with a total of 18 topic classes. With regard to the gathered social media data, the authors collected Twitter and Facebook posts from candidates and social media users. However, even tough authors gathered data from both candidates and social media users, authors split their findings depending on the used data, therefore the findings obtained

exclusively using candidates messages could be taken into account. Among their findings, the most noteworthy discoveries are:

- Politicians prefer Twitter over Facebook to comment events such as TV debates.
- Politicians use Twitter and Facebook differently. Whereas Facebook is used to mobilize users to attend campaign celebrations or similar events, Twitter is used for political debates where politicians discuss about several policies giving their own opinion. This is relevant for the proposed approach in this dissertation, since it validates our decision of choosing Twitter as the analysed Social Network.
- Politicians discuss different topics with respect to the priorities shown by electors on the surveys.

(Yaqub et al., 2017) analysed the 2016 US presidential elections' political discourse on Twitter from two points of view: studying public opinion gathering Tweets of over a million users in order to identify their talking points and behaviour(if they share original opinions, interact with other people, etc.) and, analysing the sentiment of the tweets sent by the Republican and Democrat presidential candidates. In this case, we are going to focus in the latter analysis as it has been previously mentioned. They assigned to each candidates tweets a sentiment score using a tool named SentiStregnth. Among their conclusions, the most noteworthy are that Donald Trump offered more optimistic messages than Hillary Clinton, with an average sentiment score of 0.3925 versus the negative average sentiment score of Hillary Clinton, -0.0125. They also performed a very simple analysis of the most frequently used terms by the candidates: Hillary, Donald/Trump and Vote from Hillary and Thanks, Hillary/Clinton and Great from Trump. In both cases, when a candidate was referring to the other, the sentiment average score was negative, confirming that both candidates used partisan rhetoric.

As seen during this section, most of the approaches for political analyses has been focused on traditional studies based on polarization (pro-against,

left-right, republicans-democrats) or content analysis works focused on some particular topics. However, none of the reviewed approaches have used in Social Networks a widely adopted methodology by political scientists which already have been proven to be successful when it comes to analyse political discourse on several topics and axes, as we propose in this dissertation. The most similar work is (Stier et al., 2018) with 18 categories, where Stier et al. used a categorisation schema applied exclusively in Germany for Election Study Surveys.

In conclusion, the number of categories used in both manual and automated approaches is far from the 56 categories presented in CMP. Moreover, most of the analysed works have designed their coding scheme for very specific tasks or goals, whereas CMP's categorisation schema allows the analysis in different areas with the advantage of already having annotated datasets. This categorisation schema is thoroughly explained in Chapter 3.

Breathe. Just breathe. Now reach out. What do you see?

Luke Skywalker

CHAPTER

3

Political discourse analysis using political manifestos

This chapter describes the theoretical foundations of the political discourse analysis methodology on which the research work conducted during this dissertation has been constructed. This methodology consists in annotating political manifestos to later apply content analysis techniques in order to study policy preferences of political parties according to their electoral manifestos.

The chapter is divided in four sections: Section 3.1 describes the CMPs, highlighting its importance in the political science community. Section 3.2 describes the Regional Manifestos Project (RMP), an extension of the original CMPs which goal is to measure Spanish political parties preferences when it comes to the distribution of power between lower and upper levels of government, European vs National or National vs Regional. Section 3.3 analyses the research works done so far in the automated used of political manifestos for different purposes. Finally, Section 3.4 describes the main criticisms made against the CMP approach.

3.1 Comparative Manifestos Project

The CMP is the most ambitious and accurate attempt done by political scientists to perform content analysis of parties' electoral manifestos to later derive policy positions of each political party depending on what each party claim in their manifestos.

The precursors of this methodology were the Manifesto Project, formerly known as the Manifesto Research Group (MRG), and nowadays as Comparative Manifestos Project (CMP)(Budge, 2001). In 2001, they created the Manifesto Coding Handbook(Volkens, 2002) which has evolved over the years. The handbook provides instructions to the annotators about how political parties' manifestos should be coded for later content analysis and a category scheme that indicates the set of codes available for codification. Nowadays, the category scheme for manifestos annotation consists in 56 categories (see Table 3.3) grouped into seven major policy areas(Volkens et al., 2019)): *External Relations*, *Freedom and Democracy*, *Political System*, *Economy*, *Welfare and Quality of Life* and *Social Groups*. Moreover, recently the CMP has added new subcategories for manifestos from countries which have recently transitioned or are transitioning from authoritarian regimes to a democratic system.

The annotation process is a two-step task: unitising and coding. Unitising consists in splitting each manifestos' text into quasi-sentences or coding units. Since one full sentence can contain more than one statement or message, there are some cases where a sentence has to be split into more than one quasi-sentences where each quasi-sentence contains a different message. Once the text has been unitised, a code is assigned to each of the quasi-sentences.

As it has been briefly explained in Section 1, several studies have used this methodology:

- Temporal or longitudinal analysis of how political parties' political discourse has evolved over the years (Benoit, 2009): authors analysed in particular Irish political parties' position on the issue of European integration. To do so, they focused on two categories of the CMP category scheme, *"Positive European Integration (108)"* and *"Negative European Integration (110)"*. The authors counted the total mentions of European

Integration (both positive and negative) by each Irish party per election year. Then, they studied the importance European Integration per election compared to other topics.

- Comparison between different type of manifestos (national and European level manifestos): (Brunsbach et al., 2012) compared German national and European Election manifestos held in 2009 using CMP data. Among their findings we can find that European manifestos are second class documents in Germany compared to national election manifestos. Moreover, authors found that European manifestos are set in a European context, leaving national policies in a second place.

- Analysing how peripheral parties vary their positions in some concrete topics depending on the elections: (Alonso et al., 2017) analysed Basque and British peripheral parties' manifestos in their respective 2011 and 2012 regional elections using RMP data. Among their conclusions are that peripheral dedicate most of the manifesto to regional level issues, ignoring most of the other level of competences (local, national or European), focusing on competence claims rather than nation building strategies,

The main left-right scaling method used by researchers in this field is the RILE scale (Budge and Laver, 2016). This scaling method divides a subset of the 56 categories of the CMP category schema in left or right categories (see Table 3.1). Therefore, once a political manifestos has been annotated, the RILE score is calculated taking into account the number of occurrences each of the categories in Table 3.1 have, finally obtaining a left-right score. Budge's RILE scale is the most used scale by the community, however, it is not the only one. Among the most remarkable scales are Social Liberal-Conservative scale or the scale that measures States Involvement in Economy(Benoit and Laver, 2007).

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

Table 3.1: Left and right categories according to the RILE Score.

In order to give some insights of the importance of the CMP, in table 3.2 some statistics about the project can be seen:

Countries	61
Elections	761
Political Parties	761
Manifestos	4550
Original Manifestos	2,522
Machine-Readable Documents	2,476
Scanned Documents with Codings	950
Documents with Digital Codings	1,323
Human Coded Quasi-Sentences	2,582,231
Human Coded Digital Quasi-Sentences	1,218,303
Peer-reviewed Articles and Book Chapters	432

Table 3.2: Comparative Manifestos Project's statistics

3.2 Regional Manifestos Project

Other projects have extended the original Manifestos Project annotation schema in order to be able to perform deeper analyses in some specific political topics

which have a particular importance in some countries.

For instance, the Regional Manifestos Project extended the original category scheme, introducing a new set of codes which allow the analysis of the sentences from another point of view: preferences concerning the distribution of powers between the state and lower level governments, Regional levels for instance, in the same country together with policy preferences specific to each electoral level.

In particular, they extended *centralization*, *decentralization* and *nationalism* categories in order to perform a deeper analysis of those political phenomenons. To do so, they added a new set of codes named *territorial demands* (see table 3.4) and added some new categories to the Manifestos Project category schema, increasing the number of categories from 56 to 78: positive interregional special relationships (1017), negative interregional special relationships(1027), positive representative democracy (2024), positive participatory democracy (2025), positive regional finance (3012), negative differential treatment among regions (3013), positive differential treatment among regions (3014), administration of justice (3031), management of natural resources (4111), equal treatment of immigrants (5032), welfare expansion for immigrants (5042), welfare limitations for immigrants (5051), education expansion for immigrants (5062), education limitation for immigrants (5071), promotion and protection of vernacular languages (6015), cultural links with diaspora (6016), positive bilingualism (6017), immigrants' negative impact on law and order (6051), immigrants positive (6082 and 7053) and diaspora positive (7054).

Furthermore, the dataset has a high annotators' intercoder reliability as it is proven by (Alonso et al., 2013). Authors explained the conducted methodology for manifestos annotation. As Alonso et al. concluded, when it comes to manual coding, it is impossible to reach 100% coincidence in the annotation between different coders. However, in order to reach a good level of reliability and therefore, coincidence between different coders, all coders were trained before starting coding real manifestos. Then, once the candidate had understood the coding process, a reliability test was performed. Only the coders whose results coincided at least in 85% with the correct codifications(defined

by the master codes), were selected for coding. Moreover, they compared the codification of two coders in 4 manifestos and they reached high correlations: 0.957, 0.987, 0.981 and 0.99. This differs from the manually annotated manifestos available in Manifestos' Project website where there is no information regarding annotators' intercoder reliability.

3.3 Automated use of political manifestos

The automatic codification of political manifestos and the use of this codification schema for the analysis of other types of political texts besides political manifestos is a rising research area. In the last years, there have been some authors who have worked in this field.

Most of the research works done so far has been focused on classifying text on a domain level, that is, in the 7 high-level policy domains: *External Relations*, *Freedom and Democracy*, *Political System*, *Economy*, *Welfare and Quality of Life* and *Social Groups*.

In 2016, Zirn et al.(Zirn et al., 2016) presented an approach for automated classification of political manifestos. The authors trained and validated their approach using 6 U.S. manifestos (Republican and Democrat manifestos from 2004, 2008 and 2012 elections). They only worked with the 7 policy domains. Their approach consisted in combining two different classifiers: one including only information about the sentence (bag of words representation of the sentence, domain of the preceding sentence and semantic similarity with the preceding sentence) and the second one a binary classifier which predicts if two adjacent sentences have the same code or not. Finally, Zirn et al. combine these two classifiers with information about the topic distribution in the corpus (rules representing the conditional probabilities of domain transitions). They reached the best performance combining the previously explained two classifiers and using a transition rule which indicates that consecutive sentences have the same domain label.

However, this approach can not be applied to other texts since Zirn et al. used the distribution of topics, sequences of topics and topic transitions in the manifestos as an extra feature. Unfortunately, this can not be applicable to

other type of political texts because it is necessary to have annotated texts in order to compute the distribution of topics. Furthermore, the structure of the text to be analysed can be completely different to the structure of a manifesto where sections or subsections with similar topics are discussed together and therefore topic transitions are easier to predict.

Nanni et al.(Nanni et al., 2016) used annotated political manifestos and speeches in order to analyse the speeches from the 2008, 2012 and 2016 US presidential campaigns in the 7 main domains defined by the manifestos project. The main difference between Nanni et al.'s work and our research is that first, we have analysed how fine-tuning the classifier improves the performance compared to exclusively classifying with annotated manifestos. Moreover, unlike them, we have worked with 56 CMP subcategories and we have analysed tweets instead of political speeches with two different types of context.

In 2017, Glavas et al. (Glavaš et al., 2017) proposed an approach for crosslingual topical coding of sentences from electoral manifestos using as training data, manually coded manifestos with a total of 77,500 sentences in four languages: English, French, German and Italian. Using Convolutional Neural Networks with word embeddings and inducing a joint multilingual embedding space, Glavas et al. obtained better results that monolingual classifiers in English, French and Italian. However, they achieved worse results with their multilingual classifier than a mono-lingual classifier in German. The rationale for this results is that as there are more annotated manifestos in English and German than in German and Italian, those extra training samples in English and German helped in the classification of sentences in languages with less training samples. According to Glavas et al., German did not have an improvement because they had two decades of German political manifestos covering a wider span of political issues with a high language variation. With regard to the comparison between their work and this PhD dissertation we have achieved better results in English and German using contextual data without any crosslingual training as it can be seen in Section 5.3. Glavas et al. obtained better results in French and Italian. However, we strongly believe that in order to perform a fine-grained analysis, all the research efforts

should be focused on improving subdomain classification so that the thorough analysis political scientists do using different type of scales can be performed.

When it comes to using the 56 CMP categories, to the best of our knowledge there has been only two works on this topic so far. (Subramanian et al., 2017) first and a continuation of this work a year later (Subramanian et al., 2018). These two works, based on the approach taken by Glavas et al. (Glavaš et al., 2017) of using multilingual embedding spaces in order to have of a larger training set for those minor languages with less training samples. Both research works had two classification objectives: 1) quasi-sentence's category (subdomains) , 2) the RILE score of the manifestos that is being classified/processed.

To do so, in the first work (Subramanian et al., 2017), aside from using multilingual embeddings spaces and splitting the dataset at document level, the authors built a model having two outputs, in other words, the model was trained at the same time for the multi-class classification task of predicting sentence's subdomain and for the regression task of computing manifesto's RILE score (left-right scale). They concluded that the joint-training was useful for the RILE score computation but not for the sentence level classification task. In the second work, (Subramanian et al., 2018) designed an architecture named Hierarchical bi-LSTM for Sentence and Document level modelling using more advanced NLP techniques that those used in their first work (word embeddings averaging). However, even though one of their hypotheses was that a model trained to predict the RILE score of a manifestos and the categories of the sentences at the same time could improve sentence classification's performance, authors were not able to demonstrate it. The model trained exclusively with sentences was performing better than the model where the RILE score was at the same time calculated. However, they found that as the number of training examples decreases, the model which objective is to predict sentences' categories and rile scores, starts to perform better than the model which only focus on sentence classification. When it comes to comparing their results with the results presented in this dissertation in Section 5.3, we outperform their approach in 4 out of 7 languages. Moreover, we have obtained these results in spite of not being manifestos annotation our final and only

goal. Therefore, approaches like the one proposed by Subramanian et al. are not useful for our task since first, there is no RILE score already computed for a tweet as there is for manifestos, and second, it is not our goal to study if cross lingual approaches improve classifiers performance (this has already been proven), our objective is to verify if add contextual data improves the performance of this classification task to later apply it on tweets. However, it would be interesting to combine cross-lingual classification techniques with our approach based on contextual data in order to check if both approaches are complementary or not.

A summary of the differences between the reviewed works and our approach can be seen in Table 3.6.

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

Table 3.3: Categories in seven policy domains (Volkens et al., 2019)

Code	Meaning
10	Local level
12	More authority for the local level
20	Regional Level
21	Less authority for the regional level
22	More authority for the regional level
30	National level
31	Less authority for the national level
32	More authority for the regional level
80	European level
82	More authority for the European level
90	International level
01	In favour of subsidiary principle
02	In favour of clear distinction between levels
03	In favour of shared authority between some levels
09	More than one level addressed at the same time.
00	No explicit claim for more or less authority

Table 3.4: Territorial Demands (Extracted from the Regional Manifestos Project's codebook(Volkens et al., 2019))

Code	Percentage
10	1.88%
12	0.53%
20	86.73%
21	0.01%
22	1.49%
30	3.94%
31	0.02%
32	0.01%
80	0.49%
82	0.01%
90	0.14%
01	0.08%
02	0.13%
03	2.72%
09	0.60%
00	1.21%

Table 3.5: Territorial Demands' distribution

Research	Granularity	Contextual data	Applicable contextual data outside manifestos	Application outside Manifestos
(Zirn et al., 2016)	Domain	Yes	No	None
(Nanni et al., 2016)	Domain	No	No	Yes
(Glavaš et al., 2017)	Domain	No	No	None
(Subramanian et al., 2017)	Subdomain	No	No	None
(Subramanian et al., 2018)	Subdomain	Yes	No	None
Our approach	Subdomain	Yes	Yes	Twitter

Table 3.6: Comparison between research works dealing with automated manifestos classification

3.4 Criticisms Against Manifestos' Project approach

Even though this approach for political manifestos content analysis has a wide and solid acceptance among political scientists, it has also received some criticisms.

On the one hand, (Mikhaylov et al., 2012) demonstrated after examining several annotators' intercoder reliability in two manifestos, that the coding process is highly prone to misclassification, proving the difficulty that this process has even for trained annotators. One of their main findings was the large amount of coding errors found in some specific categories. According to them, there are some coding categories that have been ambiguously defined or clearly overlaps some of the other categories. For instance, two of the most representative categories for this issue are *401 - Free enterprise* (favourable mentions of free enterprise capitalism; superiority of individual enterprise over state control systems...) and *402 - Incentives* (need for wage and tax policies to induce enterprise) where even though there is a difference between them it is easy to find a text portion were both categories could fit. Thus, these ambiguities and overlaps lead to the fact that some categories are harder than others when it comes to their assignation to quasi-sentences. Mikhaylov et al. concluded their work identifying the two most susceptible categories to coding errors: *Political Authority (305)* and *Economic Planning: Positive (404)* . However, even though this may be seen as something minor, it gains importance when we realize that this two codes are used for the RILE Score computation, being 305 a right-side category and 404 a left-side category. Therefore, a miss-codification of this categories could lead to an incorrect RILE score of a political manifesto.

To reduce the complexity of this annotation task, and therefore, facilitate to non-experts the annotation of political texts, the authors proposed, with the collaboration of the previously mentioned Mikhaylov, a new political text annotation methodology (Benoit et al., 2016). The new proposed categorisation schema's goal was to reduce the number of codification errors that may occur due to the large amount of categories the CMP codification schema has.

Their goal was to apply their methodology in order to locate parties on policy dimensions using text as data. According to them, the designed annotation methodology was reproducible (which means that the data generation process is quick, inexpensive and reliable) and agile (which means that it can be adapted depending on the needs of the specific research project).

However, unlike the CMP, their codification schema was designed having in mind two policy dimensions economic policy (right-left) and social policy (liberal-conservative).Therefore, each sentence in the document had to be annotated as a statement referring to economic policy (left-right), social policy (liberal-conservative) or to neither.

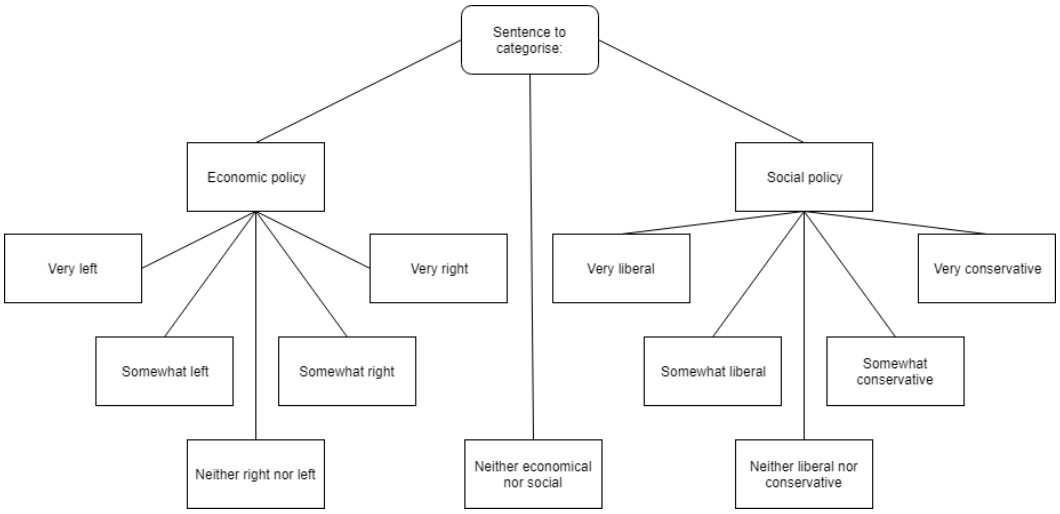

Figure 3.1: Proposed hierarchical coding scheme by Benoit et al (Benoit et al., 2016)

*People assume that time is a strict progression of cause to effect, but *actually* from a non-linear, non-subjective viewpoint - it's more like a big ball of wibbly wobbly... time-y wimey... stuff*

The 10th Doctor

CHAPTER

4

Use of the context for the design of architectures for Political discourse classification

In order to analyse how different types of contextual information could improve the political manifestos classification task, we have selected and adapted to the problem two state of the art deep learning models for text classification, modifying them for the task accomplished in this PhD dissertation.

The chapter is divided in two main sections: Section 4.1 describes the different types of contextual information we have analysed. Then, Section 4.2 explains the performed modification in the used state of the art text classification models: Convolutional Neural Networks and Bidirectional Encoder Representations from Transformers (BERT).

4.1 Analysed contextual information

As we have been introducing during the previous chapters, in this PhD dissertation we are going to analyse and validate how different contextual information could improve the automated manifestos classification tasks.

With the latest advances in NLP and text categorisation techniques, and how fast this progress is happening, analysing which architecture among all the models achieving state of the art results fits better this problem was not our objective. Therefore, we have analysed how contextual information improves this research problem with two models: Convolutional Neural Networks adapted for text classification(Kim, 2014), the state of the art model when this PhD dissertation started and BERT(Devlin et al., 2018), the state of the art model for most of the NLP task when this dissertation is being written. By doing this, we can still validate that used contextual information is still improving the performance of the latests deep learning approaches where most of the features are automatically extracted by them.

Concerning the identified different contextual data, we have worked with two types of data: the previous part of the discourse o who has said the statement in terms of political party.

The reason why the previous phrase was chosen is due to how political manifestos are annotated. During the annotation process, sentences containing more than one idea/category are divided into quasi-sentences. Therefore, it may happen that quasi-sentences of very few words without any other information are impossible to classify correctly without additional context which in this case would be the previous quasi-sentence. Moreover, this approach is also usable in Twitter where due to the character limitations of Twitter, a message sent by a politician or political party could take more than one tweet, creating a thread of tweets. Therefore, knowing the previous tweet could give some insight about what is talking about and clarify the meaning of the analysed tweet.

The second contextual data is the sender of the message. In the case of the manifestos, the sender is the political party who has written it. Conversely, on Twitter, tweets can be sent by political parties' official twitter accounts or

by politicians who are part of a political party. Therefore, we have to represent the sender of the message in a way usable in both worlds, manifestos and Twitter. Thus, even though there is some cases where politicians' language or discourse may differ, we have decided to represent each politician as its political party, supposing that most of the politicians will have a similar discourse to that of his/her political party.

However, once we have decided this new type of contextual data, a new challenge arises, how to represent the political party in the most meaningful way for the neural network. To do so, we have adopted several approaches, with different ways of representing them and using different information in order to differentiate one party from the others:

- *One hot encoding representation of each party*: one of the methods used to represent categorical variables (a list of political parties in this case) is the one hot encoding representation. This method represents each categorical variable as a list of 0s with a length equal to the number of categorical variables to represent. Then, in order to have unique representation of each variable one of the 0s is replaced by 1 and in the end, each variable will have the non zero value in one specific position which indicates which categorical variable is representing the encoding. For instance, if there were two parties, there would be an array of size 2, $[1, 0]$ representing the first party and $[0, 1]$ the second one. However, this approach has a priori two major drawbacks. First, it does not provide any information regarding parties political orientation and therefore, each party is equal to each other at the beginning of the training process even though they are diametrically opposed. Second, since the number of political parties has to be defined before training the model, every time a manifesto of a new political party wants to be added to the model, it would have to be retrained from scratch. Furthermore, manifestos of new political parties could not use this contextual information because those new parties would be unknown for the model, which derives in a scalability issue.

- *Using the average rile score of the political party*: as it has been previously introduced in Section 3.1, the RILE score is a metric that allows political scientists to place political parties and their manifestos in a left-right axis. Therefore, in order to obtain a RILE score for each political party, we have calculated the average of each political party's' manifestos' RILE score. Finally, in order to fed the neural network these new values, two feature scaling techniques have been used: standardization and normalization. The former rescales the values so that they follow a Gaussian distribution, whereas the latter, using min-max scaling, shrinks the range of the values from 0 to 1.

- *Using parties' political orientation to build a disentangled representation of the parties (see Table A.1 for a list of each analysed political parties and their corresponding political orientations)*: as it has been mentioned above, the one hot encoding approach has a priori two major drawbacks. In order to address those issues, we propose a novel approach for political parties' representation. As it is known, parties have a political orientation which gives some hints about parties' ideology and how their policies will be. Therefore, we have designed a novel method for political parties representation using their political orientation. We have extracted each political party's political orientation for European political parties from (Nordsieck, 2015), a guide with the parliamentary elections and governments since 1945 where more than 700 parties are listed with their respective political orientations, and from Wikipedia for the rest of world parties, obtaining only those orientation with references. This approach is based on the concept of disentangled representation(Bengio et al., 2013), distributed representations whose latent variables (dimensions of the vector) are semantically interpretable. In this case, a disentangled representation has been used in order to encode political parties using their political orientation. Therefore, each possible political orientation will be a dimension in the vector which represents the party and if the party follows a particular political orientation, the dimension corresponding to the orientation will be

activated in the parties' representation. For this PhD dissertation, we have identified a total of 70 different political orientations distributed among 146 different political parties from all over the world (see Table A.1 in order to see the political orientations of each party). For explanatory purposes, a small example where three parties are codified will be introduced. Assuming that there are only 6 possible political orientations, Green Politics, Euroscepticism, Right-Wing Populism, Economic Liberalism, Christian Democracy and Separatism, we want to represent the following political parties: Australian Greens (Green Politics), UK Independence Party (Euroscepticism, Right-Wing Populism, Economic Liberalism) and Basque Nationalist Party (Christian Democracy and Separatism). Each of the vectors representing the parties would have 6 dimensions (one per possible political orientation). Australian Greens would be represented as $[1,0,0,0,0,0]$, UK Independence Party $[0,1,1,1,0,0]$ and Basque Nationalist Party $[0,0,0,0,1,1]$. This allows the addition of new political parties which were not in the training process. For instance, if we want to add Team Stronach for Austria (Euroscepticism and Economic Liberalism), it would be simple, $[0,1,0,1,0,0]$. In Figures 4.1 and 4.2, the main differences between one hot and disentangled representation can be seen with the examples mentioned above.

As we have previously explained, we have considered this type of contextual information because we intuit that all this extra-knowledge could benefit the annotated tweets classification task. Our goal is first to achieve the best obtainable results for automated manifestos annotation and then, adapt all the knowledge represented in the created model to Twitter. To do so, we expect that fine-tuning with tweets and their corresponding contextual data, the model trained with annotated manifestos, could create the connection between manifestos and Tweets in order to be complementary data. For instance, in the case of who has said a statement, it could mean something similar in Manifestos and Twitter. However, the previous statement means something different in manifestos (the previous idea said in the manifesto) compared with Twitter where is the previous tweet.

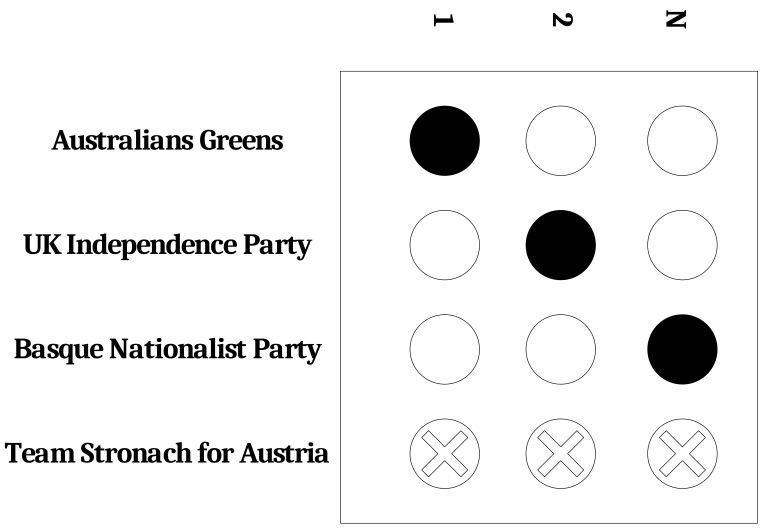

Figure 4.1: One hot representation of 3 known parties and 1 unknown party for for the model.

4.2 Text Classfication Models

4.2.1 Convolution Neural Network for Text Classification

Convolution Neural Networks (CNN) have achieved excellent results in several text classification tasks such as (Kim, 2014), (Poria et al., 2015) or (Poria et al., 2016). This, combined with the fact that this type of classifier allows the extraction of knowledge from non-annotated texts using word embeddings which later are fine-tuned to the task, has resulted in a competitive deep learning architecture.

First, the flowchart of the model will be explained as an introduction (see Figure 4.3) and after that, the model is explained in more detail. The simplified flowchart of the model is the following:

1. The phrase and the previous phrase are inserted as a list of words.
2. The embedding matrix replaces each word with its corresponding word vector, generating a sequence of word vectors from a sequence of words.
3. The phrase and the previous phrase are fed into two different structures of convolutional neural networks with 100 filters and filter sizes of $2 \times d$,

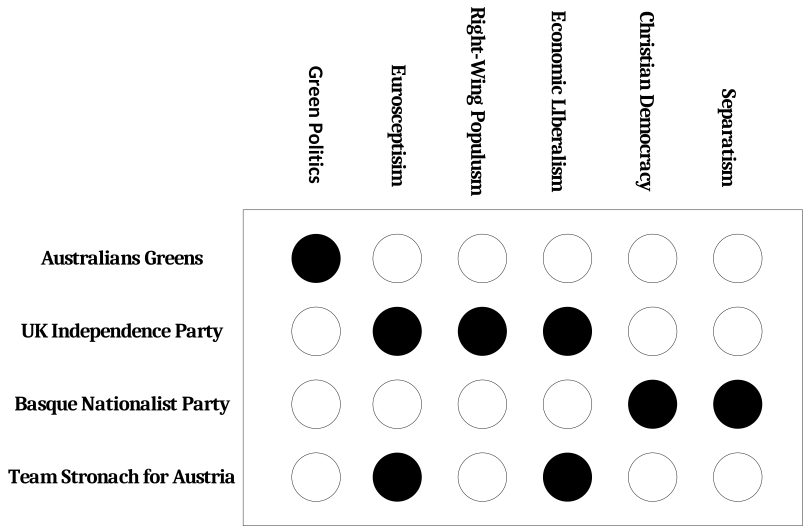

Figure 4.2: Disentangled representation of 3 known parties and 1 unknown party for for the model.

$3 \times d$ and $4 \times d$, being d the dimension of the word embedding.

4. The 1-max-pooling reduces the dimensionality of the feature maps generated by each group of filters.

5. Once their dimensionality has been reduced, the feature maps generated from the phrase and the previous phrase are concatenated.

6. If the political party to which the text belongs to is used, its representation is concatenated with the feature extracted from the CNNs.

7. A dropout rate of 0.5 is applied to the concatenation between the extracted features and the representation of the party.

8. Then, to classify the phrases to the objective political topics, a fully connected layer with ReLu as activation function is used.

9. A dropout rate of 0.5 is applied to the fully connected layer.

10. The fully connected layer with softmax as activation function computes the probability distribution over the labels.

The inputs of the model are the sentences (from manifestos or tweets) which are fed to the neural network as sequences of words. These sequences have a maximum length of 60 words. The maximum length has been decided after an analysis of the corpus' sentences' length and detecting that most of the sentences have 60 or less words.

However, the words are not provided as raw text to the convolutional neural network. The words are presented as word vectors, a multidimensional representation of each word. Those word vectors have been generated using the Word2Vec(Mikolov et al., 2013) unsupervised learning algorithm, which produces a large vector space having non-annotated raw text as input. Using Word2Vec, each word of the corpus is positioned in a multidimensional vector space taking into account its context (its surrounding words). Word's position in the N-dimensional vector space (being N the number of dimensions of the defined vector space) is used as its representation (word vector).

For example, given a sentence $S = [w_1, w_2, w_3 ... w_n]$ (n is the number of words in the sentence), the context of the word w_i would be $Context_k(w_i) = [w_{i-k}, ..., w_{i-1}, w_{i+1}, ..., w_{i+k}]$ where $2k$ is the window size for the context. Then, the log-likelihood is maximized in order to compute the word vector of each word:

$$J_{ML} = logp(w_i | Context_k(w_i))$$

300 has been chosen as word vectors' size (number of dimensions of the multidimensional space where the words are positioned) to take advantage of already pre-trained Word2Vec models in several languages published by Kyubyong Park (Park, 2018). However, for the Spanish Word2Vec model a different pre-trained model has been used(Almeida and Bilbao, 2018), created with a corpus of 3 billion words. In the case of the English Word2vec, we have used a Word2vec model pretrained with Google News corpus (3 billion running words).

Once all the word vectors have been computed, the following operation is performed. First of all, a dictionary D where words are mapped to indexes $(1, ..., |D|)$ is created, being $|D|$ the number of unique words in the corpus and

saving the 0 index for padding purposes. Therefore, the input sequences of words are transformed into a sequences of 60 indexes, padding with 0s those phrases which have a length of less than 60 words, since CNNs does no admit different sizes for the input data once the input size has been set. Then, these indexes are transformed into their corresponding word vector using an embedding layer or matrix. This embedding matrix acts as a dictionary: having the word index, the embedding matrix returns the corresponding word vector which has been previously computed. The embedding matrix is generated concatenating all the vector representations of all the existing words in D, creating a matrix $W \in \mathbb{R}^{|D| \times d}$, where d represents the vector size of the word embeddings which is 300 in this research.

Therefore, the embedding matrix works as a dictionary whose input is the word index and its output is the vector representation as it can be seen in Figure 4.4. The embedding matrix can be both static or non-static. On one hand, the static approach treats all the word vectors as static values which can not change through the training process and therefore all those weights per word defined by Word2Vec remain constant through all the training. On the other hand, a non-static embedding matrix changes as the training process evolves since the word vectors are interpreted as new parameters for the model and they are fine-tuned during the training. Non-static word-embedding have been used since it improves the model's performance(Kim, 2014).

Once the phrase has been transformed from a sequence of words to a sequence of word indexes and finally to a sequence of word vectors (see Figure 4.4), the phrase can finally be fed into the convolutional neural network, since the sequence of word vectors are in fact a matrix which dimensions are $60 \times d$ where convolution operations can be performed.

The Convolutional Neural Network (CNN) are a specific type of neural networks with neurons, weights and biases where convolution operations are performed and have been traditionally used for recognizing visual patters directly from images (pixels)(LeCun et al., 2015). However, as previously has been explained, in recent years, CNNs has also been used for text classification. In brief, convolution operations consist in moving different windows (filters made of neurons) with different sizes, s (filter sizes) analysing different

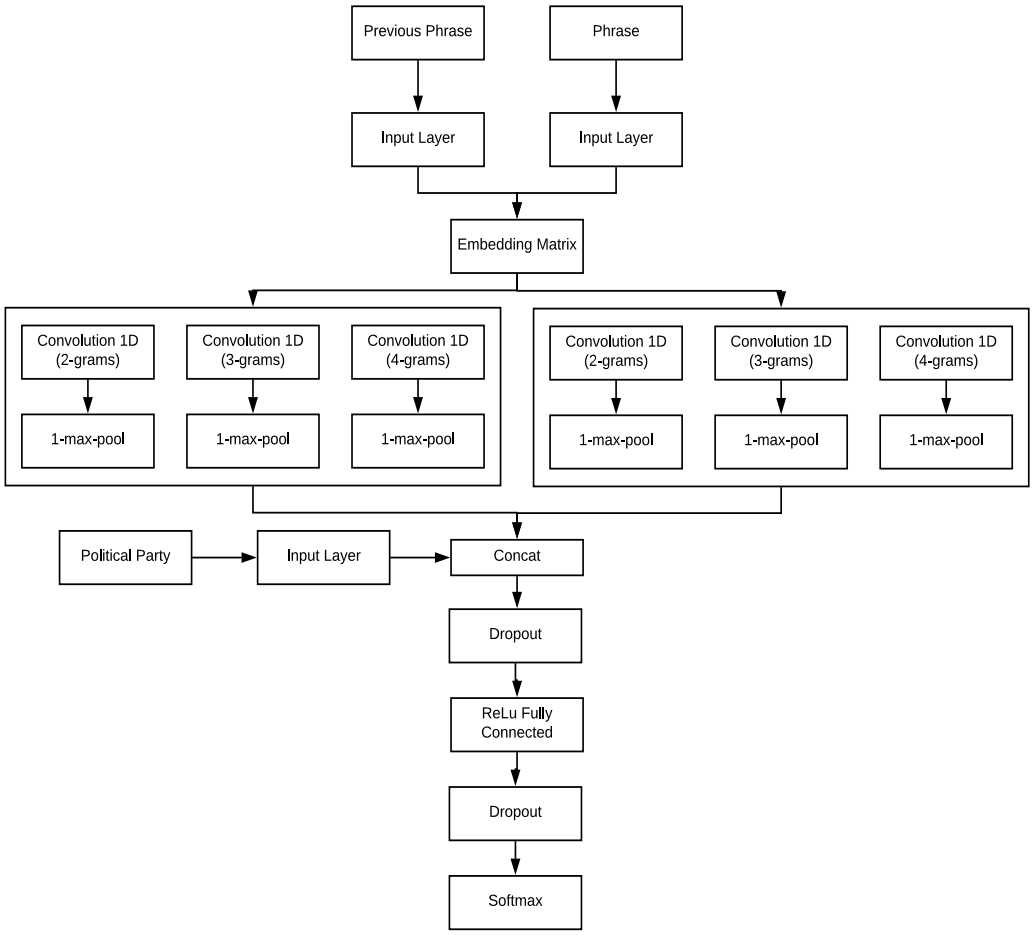

Figure 4.3: Multi-scale CNN architecture for political discourse analysis

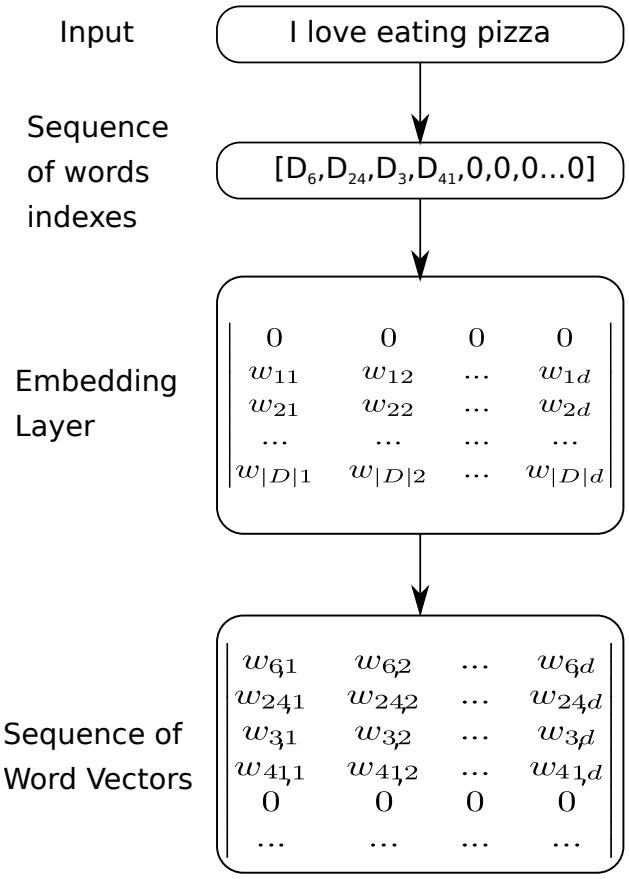

Figure 4.4: Raw text transformation into a matrix of word vectors.

regions in the matrix (an image or a list of word vectors) to extract different features. The proposed model performs convolution operations with 3 different filter sizes, batch normalization(Ioffe and Szegedy, 2015) and ReLU as the activation function. Batch normalization acts as an extra regularizer and increases the performance of the model.

The defined filter sizes are $2 \times d$, $3 \times d$ and $4 \times d$. These filter sizes can be compared to a selection of n-grams: bigrams, trigrams and fourgrams respectively. As it can be seen in Figure 4.4, each row in the matrix represents a word and therefore a filter size of $2 \times d$ will take the whole width of all the possible bigrams of the sentence, filter size of $3 \times d$ all the possible trigrams and filter size of $4 \times d$ all the possible fourgrams. This is how a single filter would work, however, as it is stated in (Zhang and Wallace, 2015), multiple filters should be used in order to learn complementary features. The model has 100 filters per different filter size. Once a filter has been applied, a feature map is generated. Therefore, a different feature map is generated per applied filter.

Then, the following operation is performed once per filter, being the filter size fs, embeddings dimensionality d and phrases length p, the input sentence is the matrix $S \in \Re^{p \times d}$. Thus, the convolution can be represented as:

$$O_j = f(W_j \circ [_1, ..., s_{p-fs+1}] + b) \tag{4.1}$$

$O_j \in \Re^{p-fs+1}$ is the result of the convolution. W_j and b are the parameters that are being trained. $f()$ is the activation function for the convolution, which in our case is a ReLU activation (Glorot et al., 2011). Finally, $W \circ S$ represents the element-wise multiplication of the elements. Being the number of filter maps d_o, the output of the convolution is $O = [O_1, ..., O_{d_o}] \in \Re^{(p-fs+1) \times d_o}$.

After the convolutional layer, there is a pooling layer whose objective is to reduce the dimensionality of the incoming data. There are different pooling strategies: average pooling, max-pooling, 1-max-pooling, etc. We have opted for the 1-max-pooling(Boureau et al., 2010) strategy since it has been proved in (Zhang and Wallace, 2015) that is the best approach for natural language processing tasks. It captures the most important feature (the highest value)

from each of the feature maps. Therefore, the output of the pooling is a feature per filter which are later concatenated into a feature vector.

Next, a dropout(Srivastava et al., 2014) rate of 0.5 is applied as regularization in order to prevent the network from over-fitting, followed by a fully connected layer with ReLU as the activation function and batch normalization. Then a 0.5 dropout is applied(Zhang and Wallace, 2015). Finally, the softmax function computes the probability distribution over the labels.

The categorical cross-entropy loss has been used as training objective function since it supports multiclass classifications. The optimization has been performed using Adam(Kingma and Ba, 2014) with the parameters of the original manuscript.

4.2.1.1 **Adding Contextual Information**

Regarding how the previous phrase has been added to the model as a new input in order to improve the performance of the model as it will be demonstrated in the next section, two different approaches have been tested:

- As a second channel in the convolutional layers: when convolution operations are applied to text only one channel is used, the channel where the sentence to be classified is inserted. However, in this approach a second channel is used to insert the previous sentence. Therefore, the convolution operations are applied to two channels.

- Replicating for the previous phrase, the same convolution-pooling process it is used in the actual phrase as it can be seen in Figure 4.3.

With regard to the political leaning, it is represented with a one-hot encoding scheme, a disentangled representation of the party or its RILE score as float. Therefore, in case of the one-hot encoding the input size of this value will vary depending on the number of political parties whose manifestos has been used to train the model. Since RILE score is a value, its size will be 1. Disentangled representations depends on the number of defined political orientations, in this case 70 (see Table A.1). Then, parties' representation has to be concatenated to the feature maps obtained after the convolutions as it can

be seen in figure 4.3. The evaluation of this approach is performed in Section 5.3 (for manifestos) and Section 5.4 (with Tweets). The results achieved with manifestos are shown in Tables 5.5 and 5.6, whereas the scores obtained with Tweets are shown in Tables 5.13 and 5.15.

4.2.2 **BERT**

Bidirectional Encoder Representations from Tansformers (BERT) proposed by (Devlin et al., 2018) has meant a considerable improvement on the NLP field. When BERT was presented, this new NLP model achieved state of the art results on eleven NLP tasks without the need to make substantial changes to the architecture. Moreover, BERT has been the first really successful attempt of transfer learning in NLP, a technique that had been successfully applied on other tasks such as computer vision but similar performances had not being achieved for NLP problems.

In particular, BERT is a pre-trained language model (LM). LMs have already shown their effectiveness on improving other model's performance in several NLP task as it is stated in (Devlin et al., 2018). There are two approaches when it comes to using pre-trained language representations or models in other tasks: feature based and fine-tuning. Feature based approaches use architectures specifically designed for the task including pre-trained representations as features, whereas fine-tuning approaches have very few task specific parameters.

However, according to Devlin et al. both approaches share a major limitation: most of the models are unidirectional (GPT (Radford et al., 2018)) or use shallow concatenations of left to right and right to left unidirectional language models such as ELMO (Peters et al., 2018). The authors give as an example of this phenomenon OpenAI's GPT (Radford et al., 2018), where each token is only able to see the previous tokens in the self attention layers of the used Transformer (Vaswani et al., 2017). However, according to them it is necessary to use context from both directions as BERT does.

In order to pre-train BERT the authors used two corpora: the BookCorpus(800M words)(Zhu et al., 2015) and the English Wikipedia (2,5000M

words). Both corpora contain text at document level, something essential according to the authors, since sentence level corpora does not have the same performance as these type of corpus, since they do not provide the necessary long contiguous sentences.

BERT is pre-trained using two unsupervised tasks with the previously mentioned datasets: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP).

MLM consists in randomly masking 15% of the input tokens to later predict the masked word using the left and right context. Once a predefined percentage of the input tokens are masked, the model (BERT in this case) is trained to predict which word the *[MASK]* token is replacing. Therefore, in this case the final layer of the model is a softmax of the same size as the vocabulary where a vocabulary ID corresponding to the replaced word is predicted.

However, the masking technique can only be applied in the pre-training process, when the model is fine-tuned words are not masked because otherwise, in other tasks, such as sentence classification or machine translation, valuable information would be lost if some word were replaced by the masking token. Therefore, in order to ease the this issue, the words chosen to be masked are not always replaced by *[MASK]*. 80% of the time are replaced by the *[MASK]* token, 10% by a random token and 10% the token remains unchanged maintaining the original token.

NSP consists in training the model to understand the relationship between two sentences for tasks such as Question Answering (QA) and Natural Language Inference (NLI). To do so, they created a corpus for a next sentence prediction task. The corpus was created randomly choosing sentences (A), being each sample a training instance, and assigning to each A, a B sentence which 50% of the times was the true next sentence and 50% a random sentence extracted from the corpus. This pre-training task has been demonstrated to be beneficial por QA and NLI tasks.

With regard to how the input sentences are processed, BERT uses a tokenization technique called WordPiece Model (WPM) (Wu et al., 2016). This segmentation technique was designed in order to tokenize input sentences in

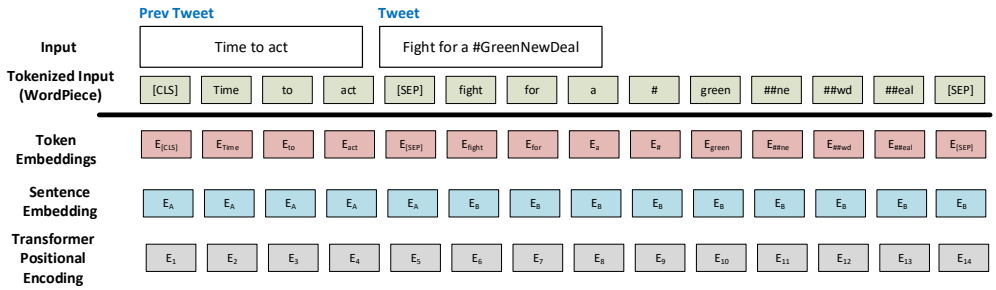

Figure 4.5: Example of how a tweet and it previous tweet would be fed to BERT. Based on the figure shown in (Devlin et al., 2018)

a deterministic way dealing with out of vocabulary words. To do so, words are divided into wordpieces or subwords that can be reverted to their original form using reserved boundary symbols. In this manner, unknown words can be decomposed into known subwords and some knowledge can be extracted from them. In particular, BERT has a 30,000 token vocabulary with some reserved special tokens such as [*CLS*] which is always the first token of each sequence or [*SEP*] to divide sentence pairs.

However, once the input sentence has been tokenized, for each of the given tokens an input representation must be built. This new token representation is constructed adding three different embeddings as it can be seen in Figure 4.5. The token embedding represents the semantic meaning of the token on a multidimensional space; the sentence embedding indicates if the token belongs to the first sentence or to the second; finally, transformer positional encoding indicates the order of the token inside the sequence of tokens.

Therefore, the first part of BERT's architecture that should be explained is the Transformer, its core module. The transformer is based on the use of self-attention for training and modelling of sequences (machine translation, language generation, etc.) without using recurrent models such as RNNs or LSTMs. To do so, they use an encoder-decoder architecture as it can be seen in 4.6. However, BERT only uses the encoder side of the transfomer. Just as BERT, the transformer encoder's needs positional encodings in order to know the place of each token in the sequence since no recurrence of any kind is used. Apart from some normalization and feed forward layers, there

is a structure named Multi Head Attention inside the encoder which is the most important element of the encoder. Each of this Multi-Head attentions implement an attention technique named *Scaled-Dot Product Attention*. This attention method consist in three inputs matrices: queries (Q) and keys (K) of dimension d_K and values of dimension d_v. Then, this attention mechanism is replicated N times (or N heads) in order to learn different features from each attention mechanism.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

(Devlin et al., 2018) built two different BERT model sizes. BERT-Base was built in order to be comparable to other approaches in the state of the art in term of parameters (12 stacked encoders, 12 self attention heads and 110M parameters); and, BERT-Large to obtain state of the start results (24 stacked encoders, 16 self attention heads and 340M parameters).

In this Dissertation we have used the BERT-Base model due to technical limitations. However, even though we have used the BASE model we have not been able to fine-tune all the model as (Devlin et al., 2018) recommends in their manuscript due to again, hardware limitations. Therefore, all the results given in Chapter 5 have been computed with all the layers frozen except the last (12th) encoder which is finetuned.

Finally, regarding how BERT is converted into a model to solve multiclass classification problems, a softmax function has to be added at the end of the 12th encoder whose output is $(128, 768)$, where 128 is the number of words and 768 is the size of the hidden state which represents each word. These values are predefined by BERT-BASE. As is in the case of CNNs, categorical cross-entropy loss has been used as training objective function and Adam as optimizer. Conversely, in order to adapt BERT to a multi-label classification problem a sigmoid function have been used instead of softmax, and binary crossentropy as loss function.

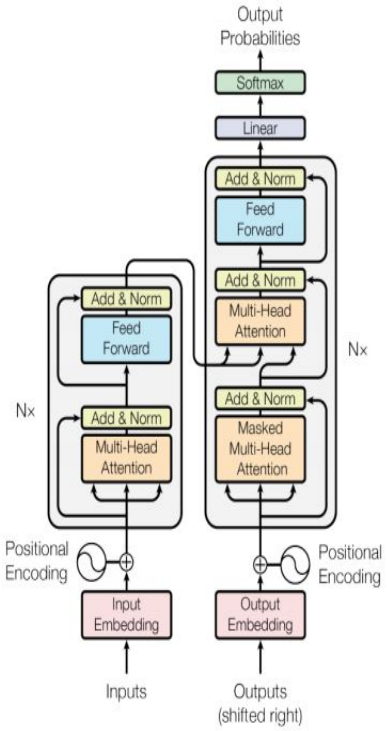

Figure 4.6: The Transformer. Encoder (left), Decoder (right). Figure obtained from (Vaswani et al., 2017) only for explanatory purposes.

4.2.2.1 **Adding Contextual Information**

With regard to how the previous phrase has been added to the model, we have taken advantage of BERT's design. Bert allows an input pair of sentences and using the sentence embedding shown in 4.5, is able to differentiate between the introduced pair of sentences. Also, BERT adds the [SEP] token in order to distinguish the sentences.

Regarding the political leaning, in this case, the representation of the party is concatenated to the parameters coming from the last encoder. Once the output of the encoder and the representation is concatenated, the probability distribution over the target labels is computed using softmax for multiclass classifications and sigmoid for multilabel. The evaluation of this approach is performed in Section 5.3 (for manifestos) and Section 5.4 (with tweets). The results for manifestos are available in Tables 5.9 and 5.10, whereas the results achieved can be seen in Tables 5.14 and 5.16.

Lights will guide you home.

Coldplay - Fix You

CHAPTER

5

Evaluation

The following chapter presents the evaluation of the work done during this dissertation. First, a brief summary of what experiments have been performed and how have been evaluated is shown in Section 5.1. After, in Section 5.2 the followed evaluation methodology is thoroughly explained. Section 5.3 shows the achieved results in the automated manifestos annotation task and compares them with other approaches in the literature. Then, Section 5.4 presents the obtained results when classifying political tweets coming from politicians' and political parties' twitter accounts. Finally, Section 5.5 shows a use case where the 2016 United States presidential elections in Twitter are analysed.

5.1 Introduction to the evaluation

The evaluation process has been divided in two different task with a final use case. First of all, we have evaluated our approach for the classification of political manifestos. As it has been stated previously, this is a multi-class classification problem (out of N classes, one class has to be selected). The results has been compared with rest of approaches followed in the literature.

Then, the evaluation with annotated tweets has been made. Similar to manifestos, first of all we have evaluated the performance of our model addressing the problem as a multi-class classification task. However, as a tweet may contain more than one idea, we have also evaluated the tweets' task as a multi-label classification problem, where out of N classes, at least one class has to be selected, in other words, more than one class can be assigned to a tweet.

Finally, we introduce a use case scenario where we show with an example how this approach could be used for political discourse analysis in Twitter.

5.2 Evaluation Methodology

Due to the imbalanceness of the datasets and a large number of categories, the results have been presented using three different measures: accuracy rate, F-Measure (Macro) and G-Mean.

Even though at first glance we tend to use the percentage of correctly classified elements (accuracy) in order to evaluate the quality of a classifier, this can lead to errors when a classifier's performance is evaluated. This issue occurs because if most of the elements belong to a specific class, a naive classifier could always classify instances with the most frequent label in the dataset and therefore achieve a high accuracy. For instance, imagine a binary classifier (two possible outcome classes) that has to classify a dataset with 10000 elements which 9000 of them are of class A. A very simple classifier could always classify the elements as class A and get an accuracy of 90% even though it is not performing a classifying task. However, the model would not work correctly with other dataset that can be balanced or even has a majority of class B elements. Therefore as it stated by (Valverde-Albacete et al., 2013), other metrics than the accuracy rate should be presented when the evaluation of this type of problems is performed. As is the case in this dissertation, where all the categories have the same priority or weight, and therefore, there is not interest in performing better with some particular categories.

In this type of problems is where metrics such as F-Measure and G-Mean should be used. F-Measure is the weighted average of precision and recall.

Precision is the number of correctly classified positive examples divided by the number of examples labelled by the classifier as positive. Recall is the number of correctly classified positive examples divided by the number of positive examples in the data and F-Measure is a combination of the two previous measures. Its objective is to consider both precision and recall measures in one unique value in order to have a single value for measuring classifiers. F-Measure reaches its best value at 1 and worst at 0. Even though the traditional F-Measure gives the same weight to recall and precision, there are some other variants of the F-Measure where recall is more important than precision or precision is more important than recall. These versions of the F-Measure are used when a problem must prioritise recall over precision or vice versa.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

However, the best way to understand what is the recall's and precision's purpose is by using an example. Imagine, for example, a classifier whose goal is to classify if a film has won an Oscar or not. The classifier is tested with 100 films, half of which have won an Oscar.

Imagine now that the classifier has classified as Oscar winners 20 out of 100 films and that those 20 selected films are definitely Oscar winners. This classifier would have a very high precision because when it says that a film is an Oscar winner, it is an Oscar winner. However, it would have a low recall because it would have detected only 20 Oscar winners out 50. The same happens to recall, imagine that the classifier has classified 80 films as Oscar winner, its recall would be very high but its precision, very low.

Therefore, the ideal classifier should have both high precision and high recall, and consequently, it would select all Oscar winners without including any film which is not. However, it is not common to achieve both high precision

and high recall values. Thus, both values are used in order to measure the performance of a classifier. F-Measure combines both recall and precision to have a single number to describe the performance of the classifier computing the mean of the precision and recall.

Unfortunately, the previously explained performance measures are focused on binary classification problems, that is to say, on problems which goal is to classify an element in two classes (yes or no, honest or dishonest, etc.). However, our classifier is a multiclass-classifier and therefore a single recall or precision measure does not really represent the performance of the whole classifier, it would only represent the performance of the classifier predicting a precise class.

Nevertheless, there are some alternatives called micro and macro averaged evaluation measures that are being used by researches in order to evaluate their multiclass-classifiers (Sokolova and Lapalme, 2009). Averaged evaluation measures consist in calculating each class' recall, precision or F-Measure and then doing their averages. However, there are two different approaches to the calculation of the averages: micro and macro averaging.

The difference between them is that macro-averaging gives equal weight to each class (in our example, each case would have 50% of the weight) when micro-averaging gives the same weight to each per-element classification, in other words, it adds all the true positives, false positives and false negatives as if they belonged to the same class.

As it is explained in (Van Asch, 2013), F-Measure ignores true negatives, and its value is basically determined by the total of true positives. Therefore, in micro-averaged measures, large classes have more influence in the averaged value than the small classes because the true negatives are not taken into account.

Label	**tp**	**fp**	**fn**	**Precision**	**Recall**
c1	10	10	10	0.5	0.5
c2	90	10	10	0.9	0.9

Table 5.1: Precision and recall averages of two classes

Following table 5.1, we know that class 2 have more elements than class 1 and their precision are 0.5 and 0.9 respectively. After that, if we calculate their micro-averaged precision $(90+10/100+20) = 0.83$, we can clearly see that the output value is closer to the largest class' precision (0.9) than to the precision of c1(0.5).

In macro averaged measures however, the mean is computed, and the smaller classes have more influence in the final value than in the case of the micro averaged measures: $\overline{(Precision(c1) + Precision(c2))} = 0.7$

Moreover, it is noteworthy to mention that even though in the other research papers which have worked with the automation of the annotation of political manifestos have been exclusively reporting their results with Micro-Averaged F-Measure, in this PhD dissertation three different evaluation measures have been used as it has been explained previously.

The rationale for not reporting our results exclusively on Micro-Averaged F-measure is due to the fact that in a multi-class problem, a Micro-Averaged F-measure will always return the same result as accuracy (% of correct samples) as it is mentioned on Scikit-Learn's documentation(Scikit-Learn-Developers): *Note that if all labels are included, "micro"-averaging in a multiclass setting will produce precision, recall and that are all identical to accuracy.* This, combined with the fact that accuracy is not the most appropriate metric when it comes to evaluating and studying how a multiclass classification problem is performing (of course it gives some insight about the performance but combined with some other metrics such as Macro-Averaged F-measure), results in the absence of a deeper understanding of how other approaches are really performing. Therefore, all the results given in this PhD dissertation will have those three metrics in order to have a deeper understanding of the achieved metrics taking into account the large number of labels, while we are still able to compare them to the previous works in state of the art.

With regard to the G-Mean or Geometric mean score is a metric used in imbalance multiclass problems (Barandela et al., 2003). Its value can vary from 0 to 1, and we have selected the macro version as it happens with F-Measure.

5.3 Evaluating political manifestos' automated annotation

5.3.1 Experimental setup

During this experimentation process 7 different classifier configurations are tested with 7 datasets of annotated political manifestos. Each dataset consist in a group annotated political manifestos in the same language. The evaluated languages are: Spanish, Finnish, Danish, English, German, French and Italian.

In order to evaluate the proposed approach and validate that contextual information could improve classifier's performance, each dataset has been divided in 2 different subsets: training and validation sets (85%), and test set (15%). The training and validation set has been used in order to create models with 5-fold cross validation to later test their performance with the same test set. The reason why the we have split the dataset in 2 subsets and then apply cross-validation to one of them is because early stopping(Prechelt, 1998) has been used in order to stop model's training when it started to over-fit. Early stopping compares the training accuracy with the validation accuracy and after some epochs without any improvements in the validation accuracy it stops the training. However, the model may have over-fitted with respect to the validation set, therefore, a third set, the test set, is needed in order to measure the real performance of the model.

The experimentation has been done using Manifesto Project's public corpus of annotated political manifestos (Lehmann et al., 2018). In particular, political manifestos written in Spanish, Finnish, Danish, English, German, French and Italian have been used (see datasets statistics in Table 5.2).

Furthermore, since the dataset is imbalanced, we have applied stratification in order to preserve the same percentage of samples for each class. Using this approach we are able to evaluate how each class is classified since it ensures that in each of the subsets there will be a representation of each class.

Apart from reporting the metrics introduced in Section 5.2, we have applied additional statistical analyses in order to statistically evaluate the improvement given by the auxiliary or contextual information. To do so, the re-

Language	# Manifestos	# Sentences
Spanish	45	78221
Finnish	14	7872
Danish	36	7559
English	115	86500
German	78	95833
French	21	8301
Italian	15	4151

Table 5.2: Datasets' statistics(Lehmann et al., 2018)

commendations provided by Demšar(Demšar, 2006) have been followed. The author proposes the use of non-parametric statistical tests to check whether there are differences among different algorithms or not, comparing the performance of the algorithms in different datasets. In this case, 7 different datasets (one per language) and each of them with two different granularities: domain and subdomains. Specifically, Demšar concludes that the Friedman test(Friedman, 1940) with the corresponding post-hoc tests is the most suitable approach when comparing more than two classifiers over different datasets.

Unfortunately, we have only been able to apply these statistical tests to the results obtained with CNNs. Friedman test requires various datasets in order to provide reliable results. However, due to the required data, hardware and time, we could not train 7 different BERT models for each of the analysed languages. Therefore, BERT has only been tested with the English dataset. On the contrary, CNNs allows the creation of 7 different models with less resources.

The defined classifier configurations for the 7 datasets are the following (experiments C6 and C7 analyse the best method for inserting the previous sentence in CNNs. Therefore, for BERT both experiments test the same, if the previous sentence improves the performance or the classifier or not):

- *C1*: Only the sentence to be classified with no additional context.
- *C2*: The sentence plus the political party which belongs to, using one hot encoding representation of each party.

- *C3*: The sentence plus the political party which belongs to, using the normalized average rile score of the party.
- *C4*: The sentence plus the political party which belongs to, using the standardized average rile score of the party.
- *C5*: The sentence plus the political party which belongs to, using the disentangled representation of the party based on its political orientation.
- *C6*: The sentence plus the previous sentence in an additional channel on the CNNs. For BERT, the previous sentence in inserted as it is explained in Section 4.2.2.1.
- *C7*: The sentence plus the previous sentence in another CNNs structure, concatenating the features extracted by both networks. For BERT, this experiment is equal to C6.
- *C8*: The sentence, the political party to which the sentence belongs to, using one-hot-encoding representation and the previous sentence in another CNNs structure.
- *C9*: The sentence, the political party to which the sentence belongs to, using disentangled representation and the previous sentence in another CNNs structure.

These are the values of the hyper-parameters for this experiments are reported in Table 5.3 for CNNs and Table 5.4 for BERT.

5.3.2 **Results**

On one hand, Tables 5.5 and 5.6 show the results of the CNNs based models for domain and subdomain per language without any statistical analysis. On the other hand, Tables 5.9 and 5.10 present the results achieved with BERT in English without any statistical analysis.

However, when an statistical analysis (the Friedman test) of the results is made, the conclusions vary slightly. The Friedman test has been applied with

Parameter	Value
Batch Size	64
Dropout	0.5
Early stopping patience	10 epochs
Filter sizes	2,3 and 4
Number of filters	100
Size of the fully connected layers	512

Table 5.3: CNNs hyper-parameters

Parameter	Value
Batch Size	64
Dropout	0.5
Early stopping patience	10 epochs
Trainable layers	The last 10
Size of the fully connected layers	512

Table 5.4: BERT hyper-parameters

two different metrics: F-measure and G-mean. On one hand, after applying the Friedman test with the F-measures, the resulting p-value is $2.2e-16$. On the other hand, when the Friedman test is applied to G-mean, the same p-value of $2.2e-16$ is obtained.

Since the two p-values are smaller than 0.01 the null hypothesis (that all algorithms perform equally) can be rejected. Once the null hypothesis has been rejected, the corresponding post-hoc tests can be performed in order to compare the different algorithms between them and analyse which are different.

In order to compare all the algorithms pairwise, Demšar proposes the use of the Nemenyi test(Nemenyi, 1962). This post-hoc tests determines the critical difference (CD) for a significance level α. Next, if the difference between the average ranking of two algorithms is greater than the critical difference, then the null hypothesis that the algorithms perform equally is rejected.

The Nemenyi test has been performed with a significance of $\alpha = 0.05$ and two different metrics: F-measure and G-mean. In both cases, the resulting

-	Spanish	Finnish	Danish	English	German	French	Italian
C1	Acc: 65.79% F1: 61.11 G-M: 75.2	Acc: 47.03% F1: 41.61 G-M: 61	Acc: 52.2% F1: 44.49 G-M: 63.42	Acc: 64.29% F1: 60.04 G-M: 75.12	Acc: 58.01% F1: 55.78 G-M: 71.63	Acc: 57.54% F1: 52.71 G-M: 69.73	Acc: 53.04% F1: 48.62 G-M: 66.66
C2	Acc: 66.09% F1:62.12 G-M: 76.32	Acc: 47.91% F1:43.38 G-M: 62.53	Acc: 54.47% F1:46.7 G-M: 64.49	Acc: 64.63% F1:60.17 G-M: 75.09	Acc: 58.41% F1:55.89 G-M: 71.69	Acc: 58.39% F1:53.77 G-M: 70.47	Acc: 54.17% F1:48.82 G-M: 66.92
C3	Acc: 65.94% F1: 61.83 G-M: 75.97	Acc: 47.34% F1: 42.05 G-M: 61.34	Acc: 53.95% F1: 46.19 G-M: 64.51	Acc: 64.63% F1: 60.05 G-M: 74.94	Acc: 58.24% F1: 55.94 G-M: 72	Acc: 58.15% F1: 53.55 G-M: 70.26	Acc: 54.56% F1: 50.23 G-M: 67.77
C4	Acc: 66.16% F1: 61.75 G-M: 75.69	Acc: 48.17% F1: 43.61 G-M: 62.51	Acc: 55.01% F1: 46.95 G-M: 65.1	Acc: 64.22% F1:59.69 G-M:74.96	Acc: 58.51% F1: 55.98 G-M: 71.76	Acc: 58.3% F1: 53.78 G-M: 70.48	Acc: 54.24% F1: 48.81 G-M: 66.97
C5	Acc: 66.18% F1: 62.05 G-M: 76.11	Acc: 48.74% F1: 43.95 G-M: 62.86	Acc: 54.36% F1: 47.25 G-M: 65.42	Acc: 64.72% F1: 60.37 G-M: 75.36	Acc: 58.31% F1: 56.29 G-M: 72.06	Acc: 57.85% F1: 53.53 G-M: 70.50	Acc: 55.85% F1: 51.15 G-M: 68.88
C6	Acc: 71.08% F1: 67.13 G-M: 79.67	Acc: 57.42% **F1: 52.58** G-M: 69.61	Acc: 58.01% F1: 50.27 G-M: 67.83	Acc: 67.85% F1: 64.17 G-M: 78.43	Acc: 64.43% F1: 62.44 G-M: 76.48	Acc: 61.96% F1: 57.77 G-M: 73.2	Acc: 59.97% F1: 55.45 G-M: 72.13
C7	Acc: 72.18% F1:68.42 G-M: 80.41	Acc: 56.68% F1:51.41 G-M: 68.73	Acc: 57.38% F1:49.73 G-M: 67.64	Acc: 68.61% F1:64.93 G-M: 78.33	Acc: 65.7% F1:64.03 G-M: 77.54	Acc: 61.94% F1:57.71 G-M: 73.27	Acc: 60.74% F1:57.07 G-M: 72.84
C8	Acc: 72.44% F1:68.82 G-M: 80.79	Acc: 57.04% F1:51.84 G-M: 69.02	Acc: 57.99% F1:50.04 G-M: 67.47	Acc: 69.02% F1:65.32 G-M: 78.41	Acc: 65.48% F1:63.43 G-M: 77.24	Acc: 62.6% F1:58.74 G-M: 73.88	Acc: 61.06% F1:56.43 G-M: 72.55
C9	**Acc: 72.52%** **F1: 68.85** **G-M: 80.89**	**Acc: 57.45%** F1: 52.53 **G-M: 69.77**	**Acc: 58.05%** **F1: 51.12** **G-M: 68.42**	**Acc: 69.04%** **F1: 65.46** **G-M: 78.87**	**Acc: 66.12%** **F1: 64.27** **G-M: 77.75**	**Acc: 63.01%** **F1: 59.51** **G-M: 74.5**	**Acc: 62.19%** **F1: 57.24** **G-M: 73.16**

Table 5.5: Domain results for each one of the experiment configuration and datasets using CNNs. The accuracy (acc), F-Measure (F1) and G-Mean (G-M) of each experiment is shown.

-	Spanish	Finnish	Danish	English	German	French	Italian
C1	Acc: 54.16% F1: 38.33 G-M: 59.24	Acc: 30.96% F1: 18.1 G-M: 38.64	Acc: 37.28% F1: 22.33 G-M: 42.07	Acc: 50.65% F1: 35.32 G-M: 58.51	Acc: 42.71% F1: 26.66 G-M: 51.02	Acc: 46.12% F1: 29.85 G-M: 50.06	Acc: 37.46% F1: 25 G-M: 42.89
C2	Acc: 55.96% F1: 41.69 G-M: 61.64	Acc: 33.07% F1: 20.65 G-M: 41.25	Acc: 38.54% F1: 24.86 G-M: 43.52	Acc: 52.18% F1: 38.89 G-M: 61.24	Acc: 43.45% F1: 28.25 G-M: 52.15	Acc: 47.27% F1: 31.94 G-M: 51.77	Acc: 43.21% F1: 31.3 G-M: 47.28
C3	Acc: 54.88% F1: 39.63 G-M: 60.15	Acc: 30.74% F1: 18.53 G-M: 39.18	Acc: 36.81% F1: 21.77 G-M: 40.77	Acc: 50.87% F1: 35.58 G-M: 59.1	Acc:43.14% F1: 27.35 G-M: 51.66	Acc: 46.05% F1: 28.78 G-M: 48.77	Acc: 37.98% F1: 24.87 G-M: 42.54
C4	Acc: 54.85% F1: 39.26 G-M: 59.64	Acc: 30.94% F1: 18.51 G-M: 38.92	Acc: 37.64% F1: 23.66 G-M: 42.67	Acc: 51.47% F1: 36.06 G-M: 59.61	Acc: 43.4% F1: 27.82 G-M: 52.07	Acc: 45.46% F1: 26.71 G-M: 47.31	Acc: 39.23% F1: 28.33 G-M: 45.44
C5	Acc: 56.03% F1: 41.64 G-M: 61.4	Acc: 32.31% F1: 19.97 G-M: 40.58	Acc: 38.32% F1: 24.94 G-M: 43.5	Acc: 52.16% F1: 38.38 G-M: 60.81	Acc: 43.5% F1: 28.81 G-M: 52.77	Acc: 47.62% F1: 32.5 G-M: 52.46	Acc: 43.66% F1: 31.56 G-M: 47.93
C6	Acc: 60.58% F1: 44.31 G-M: 63.04	Acc: 35.56% F1: 19.99 G-M: 40.57	Acc: 42.15% F1: 24.41 G-M: 42.97	Acc: 55.35% F1: 38.74 G-M: 60.89	Acc: 48.77% F1: 32.27 G-M: 55.82	Acc: 50.26% F1: 29.33 G-M: 50.32	Acc: 45.83% F1: 30.17 G-M: 46.92
C7	Acc: 61.2% F1: 45.04 G-M: 64.27	Acc: 36.49% F1: 22.67 G-M: 43.33	Acc: 40.56% F1: 24.6 G-M: 43.3	Acc: 55.7% F1: 40.73 G-M: 63.2	Acc: 50.69% F1: 34.21 G-M: 57.87	Acc: 52.02% F1: 33.8 G-M: 53.51	Acc: 45.36% F1: 30.32 G-M: 47.21
C8	Acc: 62.31% F1: 47.7 G-M: 66.14	Acc: 39.03% F1: 24.49 G-M: 44.86	Acc: 41.15% F1: 25.55 G-M: 44.3	**Acc: 56.85%** F1: 42.73 G-M: 64.56	Acc: 50.84% **F1: 35.68** **G-M: 58.72**	**Acc: 53.72%** **F1: 38.17** **G-M: 57.04**	**Acc: 49.66%** F1: 34.66 G-M: 50.6
C9	**Acc: 62.67%** **F1: 48.58** **G-M: 66.97**	**Acc: 39.17%** **F1: 24.9** **G-M: 45.42**	**Acc: 42.53%** **F1: 27.55** **G-M: 45.99**	Acc: 56.64% **F1: 43.48** **G-M: 65.2**	Acc: **51.15%** F1: 35 G-M: 58.5	Acc: 53% F1: 37.56 G-M: 56.68	Acc: 49.34% **F1: 34.86** **G-M: 50.67**

Table 5.6: Subdomain results for each one of the experiment configuration and datasets using CNNs. The accuracy (acc), F-Measure (F1) and G-Mean (G-M) of each experiment is shown.

-	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	-	3	1	1.35	**3.42**	**4.35**	**5.21**	**6.42**	**7.35**
C2	3	-	2	1.64	0.42	1.35	2.21	**3.42**	**4.35**
C3	1	2	-	0.35	2.42	**3.35**	**4.21**	**5.42**	**6.35**
C4	1.35	1.64	0.35	-	2.07	3	**3.85**	**5.07**	**6**
C5	**3.42**	0.42	2.42	2.07	-	0.92	1.78	3	**3.92**
C6	**4.35**	1.35	**3.35**	3	0.92	-	0.85	2.07	3
C7	**5.21**	2.21	**4.21**	**3.85**	1.78	0.85	-	1.214	2.14
C8	**6.42**	**3.42**	**5.42**	**5.07**	3	2.07	1.21	-	0.92
C9	**7.35**	**4.35**	**6.35**	**6**	**3.92**	3	2.14	0.92	-

Table 5.7: Differences between the average ranking of the tested algorithms computed with the Nemenyi test ($\alpha = 0.05$) and F-measures of the classifiers.

-	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	-	2.78	0.85	1.21	**3.57**	**4.35**	**5.28**	**6.28**	**7.14**
C2	2.78	-	1.92	1.57	0.78	1.57	2.5	**3.5**	**4.35**
C3	0.85	1.92	-	0.35	2.71	**3.5**	**4.42**	**5.42**	**6.28**
C4	1.2142	1.57	0.35	-	2.35	3.14	**4.07**	**5.07**	**5.92**
C5	**3.57**	0.78	2.71	2.35	-	0.78	1.71	2.71	**3.57**
C6	**4.35**	1.57	**3.5**	3.14	0.78	-	0.92	1.92	2.78
C7	**5.28**	2.5	**4.42**	**4.07**	1.71	0.92	-	1	1.85
C8	**6.28**	**3.5**	**5.42**	**5.07**	2.71	1.92	1	-	0.85
C9	**7.14**	**4.35**	**6.28**	**5.92**	**3.57**	2.78	1.85	0.85	-

Table 5.8: Differences between the average ranking of the tested algorithms computed with the Nemenyi test ($\alpha = 0.05$) and G-means of the classifiers.

critical difference is 3.2716. Therefore, if any of the average ranking of two algorithms shown in tables 5.7 and 5.8 is greater than the critical difference, then the null hypothesis is rejected and it can be affirmed that the two algorithms have a different behaviour.

-	English - CNNs	English - BERT
C1	Acc: 64.29% F1: 60.04 G-M: 75.12	Acc: 65.9% F1: 61.47 G-M: 75.42
C2	Acc: 64.63% F1:60.17 G-M: 75.09	Acc: 66.8% F1:62.9 G-M: 76.17
C3	Acc: 64.63% F1: 60.05 G-M: 74.94	Acc: 66.31 } F1: 61.72 G-M: 75.45
C4	Acc: 64.22% F1: 59.69 G-M: 74.96	Acc: 66.51 % F1: 61.87 G-M: 75.25
C5	Acc: 64.72% F1: 60.37 G-M: 75.36	Acc: 66.36% F1:62.32 G-M: 75.92
C6	Acc: 67.85% F1: 64.17 G-M: 78.43	Acc: - F1: - G-M: -
C7	Acc: 68.61% F1:64.93 G-M: 78.33	Acc: 69.2% F1:65.4 G-M: 78.27
C8	Acc: 69.02% F1:65.32 G-M: 78.41	**Acc: 69.66%** **F1: 65.79** G-M: 78.2
C9	Acc: 69.04% F1: 65.46 **G-M: 78.87**	Acc: 69.5% F1: 65.68 G-M: 78.32

Table 5.9: Domain results for each one of the experiment configuration and model (CNNs or BERT). The accuracy (acc), F-measure (F1) and G-Mean (G-M) of each experiment is shown. C7 is not reported for BERT since it is equal to C6 as it is explained in Section 5.3.1

-	English - CNNs	English - BERT
C1	Acc: 50.65% F1: 35.32 G-M: 58.51	Acc: 53.37% F1: 40.42 G-M: 62.43
C2	Acc: 52.18% F1: 38.89 G-M: 61.24	Acc: 55.46% F1: 44.06 G-M: 65.37
C3	Acc: 50.87% F1: 35.58 G-M: 59.1	Acc: 53.1% F1: 40.18 G-M: 62.24
C4	Acc: 51.47% F1: 36.06 G-M: 59.61	Acc: 53.69% F1: 40.72 G-M: 62.84
C5	Acc: 52.16% F1: 38.38 G-M: 60.81	Acc: 55.1% F1: 43.45 G-M: 64.5
C6	Acc: 55.35% F1: 38.74 G-M: 60.89	Acc: -% F1: - G-M: -
C7	Acc: 55.7% F1: 40.73 G-M: 63.2	Acc: 57.22% F1: 44.2 G-M: 65.5
C8	Acc: 56.85% F1: 42.73 G-M: 64.56	**Acc: 58.64%** **F1: 47.1** **G-M: 67.8**
C9	Acc: 56.64% F1: 43.48 G-M: 65.2	Acc: 58.29% F1: 46.68 G-M: 67.36

Table 5.10: Subdomain results for each one of the experiment configuration and model (CNNs or BERT). The accuracy (acc), F-Measure (F1) and G-Mean (G-M) of each experiment is shown. C7 is not reported for BERT since it is equal to C6 as it is explained in Section 5.3.1

5.3.3 Discussion

During this discussion, first of all, we are going to focus on the results obtained without taking into account any statistical analysis. Those results are shown

in Tables 5.5 and 5.6 for CNNs, and Tables 5.9 and 5.10 for BERT. After that, an in-depth discussion of the results obtained with the statistical tests is presented.

With regard to the achieved raw results without any further analysis, the following conclusions can be drawn:

- Adding the previous sentence improves the performance of the classifier on domains and subdomains in both used text classification models. Regarding how the previous statement should be inserted into CNNs, there is a major improvement when the previous sentence is added duplicating CNNs structure.

- Adding the political party to which the text belongs to, significantly improves the performance of the classifier with CNNs and BERT. However this only happens using two of the four methods applied for political party representation: one-hot encoding and disentangled representation using parties' political orientation. The two other approaches for political parties representation using the RILE score achieved similar or slightly better results than the baseline without any contextual information, far from the results obtained with the other two approaches. Also, the achieved improvement is more remarkable when classifying the sentences on subdomains.

- We have not been able to detect a significant difference when comparing the one-hot encoding and disentangled representations' performance in BERT. However, the latter achieves betters results in some languages when CNNs are used. Achieving an improvement in Spanish of 3.31 points in the F-Measure of subdomains with respect to the baseline, 1.87 in Finnish, 2.61 in Danish, 3.06 in English, 2.15 in German, 2.65 in French and 6.56 in Italian.

- With regard to complementarity of the previous phrase and the political party, as the best results are obtained when both features are used, we can assume that both features are complementary. However, as it

happens when only the political party is used, the improvement is bigger on subdomains.

- As it can be seen in Table 5.9 and 5.10 where results achieved with CNNs and BERT are compared, our approach of using contextual information in order to enrich manifestos classification still improves BERT's performance, achieving around a 3 points improvement compared to using CNNs. It is also noteworthy to mention how the achieved improvement is greater for subdomains than for domains. This may have happened because we are reaching the performance peak for the domain classification task with the used dataset.

- Therefore, it can be concluded that both analysed contextual information improves the performance of the classifier in CNNs and BERT.

After analysing the results shown in Tables 5.7 and 5.8, where they are statistically studied, the following conclusions can be drawn:

- The comparisons where the null hypothesis has been rejected (in bold font in Tables 5.7 and 5.8) are equal in both tables. Therefore, the same conclusions can be reach using F-measure and G-Mean metrics.

- In both cases it is statistically validated that adding the previous phrase in an additional channel (C6) or another CNNs structure (C7) have a different behaviour than the baseline without any context data. Therefore, it can be affirmed that adding the previous phrase in an additional channel or as another CNNs structure improves the performance of the classifier, as it can be seen in Tables 5.5 and 5.6.

- It has also been statistically validated that adding the disentangled representation of the political party using their political orientation does improve the classifier's performance (C5). On the contrary, it has not been possible to statistically validate that there is an improvement when one hot encoding representation is used (C2), even though there is an improvement in performance in the majority of performed experiments.

As it has been previously mentioned, the improvement is more remarkable when classifying subdomains, therefore, the improvement in C2 could have not been statistically validated because the improvement is not enough when classifying domains.

- Even though the best results have always been obtained combining both contextual features in C8 and C9, we have not been able to statistically validate that these two configurations have a different behaviour than C6 or C7 where the previous phrase is used.

Moreover, in order to clarify the improvement given by using political party using one hot encoding, an extra statistical analysis have been made. Another Nemenyi test has been performed again with greater significance level or α values: $\alpha = 0.1$. The Nemenyi test with $\alpha = 0.1$ varies the the critical difference value from 3.2716 to 2.998914 with respect to $\alpha = 0.05$. With this new value, the difference between C1 and C2 would be higher and therefore it could be said that C2 has a different behaviour than C1, confirming that the political party represented as a one hot encoding does improve classifier's performance. This may happen due to the fact that we have not performed the test with the necessary number of datasets to confirm our hypothesis with a higher level of confidence.

So far, we have only been able to validate the improvement gained by using disentangled representations of political parties using their political orientations. However, as it has been previously explained in Section 4.1, one of the advantages this could have is the easy addition of new political parties to the designed tool, without the need of retraining the whole model and using the knowledge obtained from other political parties' manifestos. Therefore, in other to evaluate if this approach for political parties' representation is easily scalable we have performed the following experiment. We have removed 5 political parties from the English training dataset and then, we have evaluate model's performance predicting this parties manifestos without any contextual data (D1), providing the political party using one hot encoding representation (D2) and its disentangled representation (D5). As it can be seen in Table 5.11, for Congress of the People, Anti-Austerity Alliance and

Scottish National Party there is a considerable improvement from D1 to D5 in terms of accuracy and F-Measure. In particular, for Anti Austerity Alliance there is an improvement of 7 points in accuracy and 6 in F-Measure. In this case, the Geometric-Mean metric has not been reported because is computed among the 56 labels and there are some cases in these experiments where the manifestos corresponding to the party do not contain all the labels. However, the F-Measure ignores the label if there are no samples available.

-	D1	D2	D5
Congress of the People (South Africa-181420)	Acc: 51.12% F1: 35.6	Acc: 55.24% F1: 32.27	Acc: **55.68%** F1: **37.33**
Labour Party (UK-51320)	Acc: **48.8%** F1: **29.52**	Acc: 48.14% F1: 29.25	Acc: 48% F1: 28.94
Anti-Austerity Alliance (Ireland-53240)	Acc: 42.99 % F1: 21.63	Acc: 45.37% F1: 23.22	Acc: **50.04%** F1: **27.75**
Australian Greens (Australia-63110)	Acc: 48.57% F1: 22.37	Acc: 48.32% F1: 22.8	Acc: **49.68%** F1: **23.65**
Scottish National Party (UK-51902)	Acc: **43.34%** F1: 25.66	Acc: 43.23% F1: 26.42	Acc: 43.07% F1: **26.92**

Table 5.11: Comparison between one hot encoding representation and disentangled representation using political orientation for classifying manifestos of unknown parties for the trained model

In the Table 5.12, we have compared our results with other results achieved in the subdomains classification in the literature of automated manifestos classification. We have achieved better results in 4 out 7 languages: Spanish (+12.31) , English (+6.85), German (+8.84) and French (+4.72). Whereas (Subramanian et al., 2018) obtained better results in 3 out 7 languages: Finnish(+5), Danish (+2.86) and Italian (+2.34). On the one hand, our approach achieved the best results on those languages with more annotated manifestos. On the other hand, (Subramanian et al., 2018) obtained better results on those languages with less annotated manifestos using a crosslingual approach where manifestos from other languages are used as extra training data in those languages with less training samples. However, our approach is

still valid in languages with less samples, therefore we can assume that applying a crosslingual approach with contextual data could obtain the best results. However, as it has been explained in the Chapter 1.2, using crosslingual approaches is out of the scope of this Dissertation. We would also want to remind that all the reported metrics in the literature are in F-Measure(micro) which is equivalent to accuracy in multi-class classification problems, and therefore, ours are the first results reported in F-Measure(Macro) and G-Mean, the metrics we believe should be used when it comes to evaluating this task and the metrics on which we have focused our efforts.

Finally, it is important to compare the achieved performance automatically classifying manifestos with the inter-coder agreement annotators have been able to achieve for manifestos annotations. As it is compiled in (Mikhaylov et al., 2012), since the creation of CMP several experiments has been performed in order to calculate the inter-coding agreement between annotators. These experiments consisted in comparing annotators codification with a predefined gold standard. Annotators taking the test for the first time achieved a inter-coder agreement of 0.7 and 0.8 the second time. However, (Mikhaylov et al., 2012) criticised these values because they were calculated using the differences between the annotators and the gold standard and not between annotators. According to them, the real inter-coder agreement is around 0.5. In any case, these inter-coder agreements value show that in languages such as English and Spanish we are near the human level performance, and therefore it is hard to expect big improvements based on these results. Moreover, following this same approach in other languages, as the amount of annotated manifestos increases in those languages or cross-lingual classification approaches improve, we could expect near human level performances in all the languages.

5.4 Evaluating with annotated political tweets.

5.4.1 Tweets annotation methodology

In order to evaluate the performance of the proposed approach for political discourse analysis in social networks, we have annotated 5,000 tweets us-

-	(Bilbao-Jayo and Almeida, 2018)	(Subramanian et al., 2017)	(Subramanian et al., 2018)
Spanish	**62.31**	-	50
Finnish	39.03	30	**44**
Danish	41.14	35	**44**
English	**56.85**	42	50
German	**50.84**	33	46
French	**53.72**	38	49
Italian	49.66	33	**52**

Table 5.12: Comparison between 3 approaches for subdomains classification. The results are given in F-Measure(micro) which is equal to accuracy in a multi-class classification problem.

ing CMP's categorisation schema. To do so, we downloaded the last 3,000 tweets from the Twitter accounts of politicians from the United Kingdom and United States. We used two publicly available twitter-lists to gather them: *cpsan/members-of-congress*[1] and *twittergov/uk-mps*[2]. Then, we randomly selected 5,000 tweets to annotate them. It is important to note that Manifestos Project's categorisation scheme was designed to annotate each sentence's topic inside the political manifesto. However, when it comes to tweets, our goal is to classify the whole tweet in one of the CMP's categories, avoiding the categorisation of each of the sentences that a tweet can contain. Therefore, when it comes to annotating the tweet, we have selected the topic that best summarises the tweets' meaning. However, in those tweets containing more than one concept, we have added some extra categories apart from the most important one in order to analyse the feasibility of transforming this multiclass classification problem to a multi-label classification one. Also, it should be mentioned that each of the tweets has been anonymized, in other words, the annotator was not aware of who had post the tweet during the annotation process in order to avoid any bias.

[1] https://twitter.com/cspan/lists/members-of-congress
[2] https://twitter.com/twittergov/lists/uk-mps

As it happens in political manifestos, the distribution of samples over the seven domains is highly imbalanced (as it can be seen in Figure 5.1): external relations (10.61%), freedom and democracy (5.58%), political system (15.16%), economy(16.35%), welfare and quality of life(28.59%), fabric of society (15.18%) and social groups(8.52%).When it comes to the distribution of subdomains', as it can be seen in Figure 5.1, the 59.08% of the samples are divided in 10 categories, whereas the rest of the samples, 40.92% are divided in the remaining 46 categories. Therefore, the most repeated categories in a descending order are: Political Authority (305), Welfare State Expansion (504), Equality (503), Environmental Protection (501), Law and Order (605), Technology and Infrastructure (411), Labour Groups: Positive (701), Market Regulation: Positive (403) and Incentives: Positive (402).

With regard to the preprocessing of the annotated tweets, they have been preprocessed for the experiments with CNNs: removing stopwords and URLs, converting all the text to lowercase, tokenizing the sentences and maintaining hashtags and user-names. However, in the case of BERT, no preprocessing has been performed, since the used word segmentation technique, WordPiece, deals with this.

5.4.2 **Evaluation methodology**

In order to evaluate our approach we have used the two datasets previously mentioned: Manifestos Project's annotated 115 political manifestos and 5,000 annotated political tweets. Since our main goal is to analyse if annotated political manifestos with the contextual information previously introduced and tweets can work together as complementary training data for our political discourse classifier, we have divided our evaluation effort in three ways. The same test set of annotated political tweets is used for the three configurations so that the results are comparable.

- T1: Trained exclusively with annotated political manifestos and evaluated with annotated political tweets.

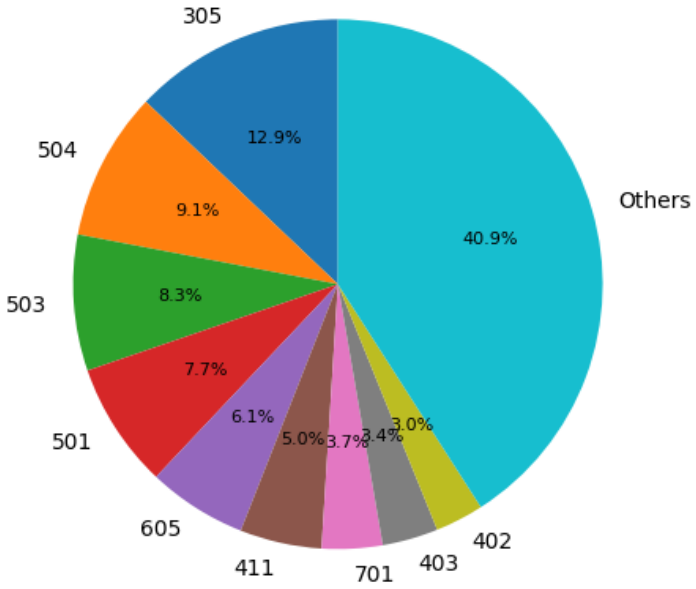

Figure 5.1: Subdomain distribution of annotated tweeets

- T2: Trained exclusively with annotated political tweets and evaluated with annotated political tweets.
- T3: Trained with annotated political manifestos and finetuned with annotated political tweets to later evaluate it with political tweets.

Therefore, the datasets has been split in the following way using stratification to maintain category distribution over all the sets:

- Annotated political manifestos: train set (75%), eval test (15%) and test set (15%).
- Annotated political tweets: train set (75%), eval test (15%) and the test set (15%) used in all the experiments.

Moreover, per each of the evaluations mentioned above, the following experiments has been conducted in order to analyse if the contextual data that

we have previously proven that does work for manifestos classification, does also work on political tweets classification:

- Analyse if the previous tweet (a tweet has a preceding tweet if the tweet is answering or quoting another one) improves the performance of the classifier.

- Analyse if the political party to which the politician posting the tweet belongs to, improves the performance of the classifier. In this case, we have tested with the two representations which had the best performances: one hot encoding and disentangled representation using parties' political orientation.

- Analyse if the previous tweet and the party responsible of the tweet ,using the two representation methods, are complementary features when are used together.

Unlike in the manifestos evaluation, in this case we have not used cross-validation in the evaluation. The reason behind this decision is the low number of annotated tweets compared to the number of samples for manifestos that would have resulted in a high variability between runs . In this experimentation, our goal has been to analyse the complementarity of manifestos datasets with respect to annotated tweets with the same codification. Therefore, in T1 the training data is the whole manifestos dataset and the test data is a subset of the annotated tweets which always will be the same during all the experiments. In T2, in order to be as comparable as possible with T1, the training data is all the annotated tweets excluding the test set. In T3, the model is first trained with the manifestos data used in T1 and then fine-tuned with the annotated tweets used in T2 for training.

As in manifestos, we have split each dataset in 3 subsets because we have used early stopping(Prechelt, 1998) in order to stop's model training as soon as start over-fitting to the train set. In T1 the evaluation set used for early stopping is a subset of manifestos, in T2 a subset of annotated tweets and in T3, first a subset of manifestos and when the model is fine-tune, a subset of

annotated tweets. Again, in all the experiments, the test set will always be the same set of tweets.

We have performed the experiments presented above with the two classification approaches used for manifestos: CNNs and BERT. Even though at first sight, after seeing the improvement achieved using BERT with respect to CNNs, it could be seen as something obvious that BERT would obtain better results than CNNs, the goal with these experiments (apart from analysing if contextual data helps), was to analyse if a language model such as BERT would perform better with tweets and without manifestos, than CNNs with tweets and manifestos. If so, this would demonstrate how powerful BERT's language model is and how good it generalises.

Also, even though we have tried avoiding the variability between executions not using cross-validation, we have found that the achieved results vary considerably among different executions. These differences were not that significant when classifying manifestos. Therefore, we have run each experiment 5 times and added to each metric its standard deviation in 5 runs.

It also should be clarified that in this case, we have not been able to perform Friedman and Nemenyi tests because we only have two datasets in order to perform then, English domain and subdomains, whereas in the case of manifestos, we had 14 different datasets and therefore the statistical test could have been applied.

5.4.3 **Discussion**

After analysing the results shown in Table 5.13 and Table 5.14 the following conclusions can be drawn when it comes to classifying tweets in the 7 high level domains of the Manifestos Project categorization scheme:

- As expected, BERT obtains better results than CNNs. In terms of F-Measure, the highest F-Measure achieved with BERT is 64.55, exclusively trained with annotated tweets and both contextual information (party with one hot encoding and previous tweet) as extra features. In addition, CNNs obtain their best F-Measure result 57.65, trained

with manifestos, fine-tuned with annotated tweets and with the previous tweet and political party using disentangled representation as extra feature. Therefore, we can affirm that at least for the classification of the tweets in 7 domains, BERT's language model fine-tuned with annotated tweets is more powerful than CNNs trained with manifestos and fine-tuned with annotated tweets.

- Even though there are much less training samples, BERT achieves better results when the model is trained exclusively with annotated tweets than with political manifestos. This may happen because first, the language used in Twitter differs from the language used in political manifestos; second, because the language used in Twitter could be can be simpler than the one used in political manifestos; and third, BERT's pre-trained language model would be enough for domain classification.

- In case of the CNNs, fine-tuning the model exclusively trained with political manifestos with annotated tweets drastically improves models performance in both accuracy and F-Measure, achieving an improvement of almost 9 points in both measures with respect to T1. On the contrary, BERT does not achieve the best results fine-tuning it with manifestos and tweets. As it has been mentioned before, the best results are achieved ignoring annotated manifestos and fine-tuning the model using annotated tweets.

- With regard to the use of contextual data, we can not affirm that the previous tweet (quoted or answering to) does contribute to an improvement when CNNs are used. However, it is true that with BERT, in those experiments where annotated tweets are part of the fine-tuning process (T2 and T3), the best results are achieved when the previous tweet is part of the used contextual data. Moreover, it is also worth mentioning that the highest standard deviation values are seen when the previous tweet is used as contextual data.

- As for the use of the political party to which the politician who has written the tweet belongs, a clear improvement can be perceived every time

this contextual data is used, both with CNNs and BERT. In fact, CNNs obtain their best results when the political party is used. In this case, one hot encoding is in most of the cases the method of representation that best works.

- With regard to the complementarity of the proposed contextual data, we can affirm that they are complementary in the experiments T2 and T3 of BERT and CNNs where the best results of this approach are achieved.

With regard to the results shown in Table 5.15 and Table 5.16 the following conclusions can be drawn when it comes to classifying tweets in the 56 subdomains of the Manifestos Project categorization scheme:

- On the contrary of what happens with Domains, in this case the best results are achieved with BERT fine-tuned with political manifestos and annotated tweets, obtaining a F-Measure of 50.07 and a G-Mean of 62.4. This values are obtained using the political party as an extra feature.

- As it happens with the high level domains, fine-tuning the model with annotated political tweets drastically improves models performance when classifying the tweets in the 56 categories. Achieving improvements in the F-Measure of more than 15 points using BERT and 10 using CNNs.

- Again, it is noteworthy mentioning the improvement gained training the model exclusively with annotated tweets (T2) compared to training it with annotated manifestos (T1). As it happens with the high level domains, this may happen due to the different language used by politicians in manifestos and Twitter. Also, this difference is significantly bigger when classifying subdomains. However, in this case the best results are achieved when annotated manifestos and tweets are combined (T3).

- Using the previous tweet as contextual data improves the performance in CNN-T2, CNN-T3, BERT-T2 and BERT-T3. This is similar to what happens when classifying high level domains, where every-time annotated tweets are used in the fine-tuning process, the previous tweet improves model's performance. This could mean that the model is not

able to adapt the meaning that the previous statement has in manifestos classification task to the meaning that the previous tweet could have when classifying Tweets. Therefore, models classifying tweets are not able to take advantage of the previous tweet/sentence until they are trained with annotated tweets, where the model is able to adapt to the new classification problem.

- The political party to which the politician who has written the tweet belongs, obtains the best results in both approaches: CNNs-T3 and BERT-T3. Regarding the method of representation, both disentangled and one-hot representation achieve similar results, being the latter the best performing in most of the cases.

- In this case, the proposed contextual data are not complementary since the best results are obtained using exclusively the political party.

Finally, we have analysed how feasible would be to change from the multiclass classification problem that we are been dealing with during this Dissertation, to a multilabel classification problem where those secondary ideas some tweets could contain are also taken into account. As it has been already explained in 5.3.1, we annotated some secondary categories (apart from the principal one), in those tweets with more than one concept. In this case, we have only used BERT as classification model and subdomains as objectives. We have used BERT because is the model that has given the best results and we have decided not to use high level domains because in most of the cases ideas inside a Tweet would belong to the same high level domain.

First, we have evaluated this task being as strict as possible, considering a corrected predicted sample a Tweet where all the labels were correctly predicted. These results are reported in Table 5.17. In this case, we have not used the G-Mean as an evaluation metric because it was not designed for multi-label evaluation. As expected since a multi label problem in this context is more complex than a multiclass problem, the results are worse than those achieved previously in the multiclass classification problem for subdomains. The best results in terms of F-Measure(Macro) has been obtained in ML2

Table 5.13: Domain results with CNNs (the average results of 5 runs per experiment are shown) for each one of the experiment configuration. The accuracy (acc), F-Measure(macro), G-mean and their respective standard deviation is shown.

Experiment	Accuracy	F-Measure	G-Mean
T1	51.57% ±0.4	47.40 ±0.5	66.93 ±0.6
T1 + Prev.Tweet	49.93% ±0.7	45.13 ±0.4	64.24 ±0.3
T1 + P.Party (One hot)	**52.16%** ±0.9	**48.03** ±0.5	**67.21** ±0.5
T1 + P.Party (Disentangled)	51.4% ±1.5	46.97 ±1.4	66.39 ±1.14
T1 + P.Party (One hot) + Prev.Tweet	50.12 ±1.3	46.24 ±1.4	65.12 ±1.1
T1 + P.Party (Disentangled) +Prev.Tweet	50.63 ±0.8	46.68 ±0.8	65.7 ±0.7
T2	57.83% ±0.4	54.62 ±1	71.21 ±0.7
T2 + Prev.Tweet	58.09% ±1.9	54.16 ±2.8	71.05 ±1.9
T2 + P.Party (One hot)	58.31% ±0.9	54.8 ±0.8	71.15 ±0.6
T2 + P.Party (Disentangled)	58.33% ±1.7	54.6 ±1.9	71.06 ±1.2
T2 + P.Party (One hot) + Prev.Tweet	**58.68%** ±0.8	**55.5** ±1.1	**71.74** ±0.7
T2 + P.Party (Disentangled) + Prev.Tweet	58.04 ±1.15	55.2 ±0.9	71.67 ±0.8
T3	59.52% ±1.8	56.2 ±2.4	71.99 ±1.6
T3 + Prev.Tweet	58.58% ±1.7	55.42 ±2.1	71.5 ±1.5
T3 + P.Party (One hot)	60.13% ±0.7	57.14 ±0.8	72.6 ±0.4
T3 + P.Party (Disentangled)	59.51% ±0.8	56.54 ±0.6	72.42 ±0.5
T3 + P.Party (One hot) + Prev.Tweet	59.73 ±1.24	57.07 ±1.36	72.96 ±1
T3 + P.Party (Disentangled) + Prev.Tweet	**60.83%** ±1.1	**57.65** ±0.9	**73.28** ±0.6

Table 5.14: Domain results with BERT (the average results of 5 runs per experiment are shown) for each one of the experiment configuration. The accuracy (acc), F-Measure(macro), G-mean and their respective standard deviation is shown.

Experiment	Accuracy	F-Measure	G-Mean
T1	55.83% ±0.8	51.39±1.25	68.65 ±1.2
T1 + Prev.Tweet	54.73% ±2.44	51.24 ±1.6	68.62 ±1
T1 + P.Party (One hot)	**57.5**% ±1.76	**53.5** ±1.88	**70.69** ±1.25
T1 + P.Party (Disentangled)	56.21% ±2.5	52.44 ±2.7	69.65 ±1.9
T1 + P.Party (One hot) + Prev.Tweet	56.88% ±1.7	52.74±1.7	69.82 ±1.3
T1 + P.Party (Disentangled) + Prev.Tweet	57.15% ±0.8	52.43 ±1	69.94 ±0.9
T2	66.79% ±1.9	63.48±2	76.7 ±1.5
T2 + Prev.Tweet	67.73% ±1-85	63.73±2.17	77.09 ±1.5
T2 + P.Party (One hot)	67.08% ±0.75	64.19 ±0.4	**77.49** ±0.7
T2 + P.Party (Disentangled)	67.22% ±1.2	63.84 ±1.1	77.23 ±0.8
T2 + P.Party (One hot) + Prev.Tweet	**67.91**% ±1.7	**64.55**±1.97	77.4 ±1.4
T2 + P.Party (Disentangled) + Prev.Tweet	67.19% ±0.7	63.84 ±0.85	77.26 ±0.8
T3	65.36% ±1.18	61.88 ±1.82	75.78 ±1.2
T3 + Prev.Tweet	66.79% ±0.9	63.44 ±1.39	76.99 ±1.14
T3 + P.Party (One hot)	65.77% ±0.7	62.23 ±2.5	76.11 ±1.7
T3 + P.Party (Disentangled)	67.16% ±0.9	63.6 ±1.4	77.02 ±0.9
T3 + P.Party (One hot) + Prev.Tweet	**67%** ±0.7	**63.57** ±1.05	**77.04** ±0.75
T3 + P.Party (Disentangled) + Prev.Tweet	66.95% ±0.7	62.75 ±1.7	76.39 ±1.1

Table 5.15: Subdomain results with CNNs (the average results of 5 runs per experiment are shown) for each one of the experiment configuration. The accuracy (acc), F-Measure(macro), G-mean and their respective standard deviation is shown.

Experiment	Accuracy	F-Measure	G-Mean
T1	37.20% ±2.1	22.83 ±1	47.68 ±1.29
T1 + Prev.Tweet	34.74% ±2.3	21.79 ±2.1	45.68 ±2
T1 + P.Party (One hot)	35.96% ±1.2	23.28 ±1.8	47.66 ±1.26
T1 + P.Party (Disentangled)	**37.57%** ±2.6	**24.14** ±2.3	**48.4** ±1.7
T1 + P.Party (One hot) + Prev.Tweet	37.14 ±1.9	23.64 ±1.4	47.83±1.46
T1 + P.Party (Disentangled) + Prev.Tweet	37.57% ±2.6	22.54 ±2.1	46.46 ±2.3
T2	42.88% ±1.88	27.58 ±1.94	45.8 ±1.33
T2 + Prev.Tweet	43.93% ±1.4	28.67 ±2.22	47.4 ±1.88
T2 + P.Party (One hot)	44.14% ±1.75	28.27 ±1.6	46.67 ±1.18
T2 + P.Party (Disentangled)	**44.66%** ±1.21	**30** ±1.75	**47.92** ±1.22
T2 + P.Party (One hot) + Prev.Tweet	42.58% ±2.2	28.21 ±2.1	47.68 ±1.3
T2 + P.Party (Disentangled) + Prev.Tweet	44.2% ±0.5	28.32 ±1.3	46.82 ±1.07
T3	49.46% ±1.4	38.06 ±2.83	54.75 ±2.1
T3 + Prev.Tweet	50.32% ±0.7	38.58 ±2.87	54.8 ±2.48
T3 + P.Party (One hot)	49.76% ±1.6	39.20 ±1.96	55.23 ±1.58
T3 + P.Party (Disentangled)	**51.99%** ±1.05	**41.43** ±1.35	**57.03** ±1
T3 + P.Party (One hot) + Prev.Tweet	50% ±2.5	37.75 ±3.4	54.64 ±2.6
T3 + P.Party (Disentangled) + Prev.Tweet	50.18% ±1.89	39.15 ±3.96	55.82 ±3.23

Table 5.16: Subdomain results with BERT (the average results of 5 runs per experiment are shown) for each one of the experiment configuration. The accuracy (acc), F-Measure(macro), G-mean and their respective standard deviation is shown.

Experiment	Accuracy	F-Measure	G-Mean
T1	44.03 % ±1.8	28.29±2	50.53 ±1.5
T1 + Prev.Tweet	43.61% ±1.79	27.81 ±0.7	50.48 ±0.7
T1 + P.Party (One hot)	45.25% ±1.47	27.8 ±0.9	51.49 ±1.1
T1 + P.Party (Disentangled)	45.71% ±1.83	**29.08** ±2.09	**52.13** ±1.82
T1 + P.Party (One hot) + Prev.Tweet	**45.95%** ±1.84	28.19±1	51.55 ±1.2
T1 + P.Party (Disentangled) + Prev.Tweet	45.57% ±1.95	28.71 ±2.26	50.87 ±2.18
T2	59.16% ±1.1	45.86 ±2	59.98 ±1.06
T2 + Prev.Tweet	59.73% ±1.7	47.01 ±2.5	60.51 ±1.9
T2 + P.Party (One hot)	58.51% ±1.89	44.06 ±2.3	58.76 ±1.4
T2 + P.Party (Disentangled)	**60.24%** ±0.6	46.66 ±1.6	60.21 ±0.5
T2 + P.Party (One hot) + Prev.Tweet	59.16% ±1.3	47.51 ±2.47	**60.89** ±1.78
T2 + P.Party (Disentangled) + Prev.Tweet	60.21% ±1.07	**47.74** ±2.6	60.57 ±1.48
T3	60.67% ±1.3	48.02±2.56	60.57 ±1.75
T3 + Prev.Tweet	60.48% ±1.8	48.19 ±2.38	60.64 ±1.64
T3 + P.Party (One hot)	60.51% ±1.08	**50.07**±3.1	62.4±2.58
T3 + P.Party (Disentangled)	**61.64%** ±1.05	49.81 ±2.66	61.83 ±1.44
T3 + P.Party (One hot) + Prev.Tweet	60.29% ±1.2	48.68 ±2.5	60.96±1.8
T3 + P.Party (Disentangled) + Prev.Tweet	60.08% ±0.6	49.94 ±1.09	**62.44** ±1.13

and ML3 using all contextual information as extra features, 42.34 and 41.71 respectively. However, ML2 without any contextual information achieves the best accuracy rate (+0.44%) but it has worse F-Measure, -0.92. Regarding the complementarity of annotated manifestos and tweets in this task, even though the best results in terms of F-Measure is achieved in ML3 (manifestos + tweets), the difference with respect to ML2 (only tweets) is minimal: +0.37. Therefore, we can not conclude that in this case both datasets are complementary.

Second, we have evaluated the task being less strict, considering a corrected predicted sample a Tweet where at least one of the labels were correctly predicted. These results are reported in Table 5.18. Predictably, the results have improved with respect to strict evaluation shown in Table 5.17. ML2 achieves its best results using the preceding tweet, 3 points better in F-Measure compared with the baseline. Also, disentangled representation for political parties improves baseline's performance and outperforms by a wide margin the one-hot econding representation. However, in this case previous tweet and political party are not complementary data. With regard to ML3, it achieves the best results, confirming the fact that annotated manifestos and tweets are complementary. In this case, previous tweet and political party are complementary data.

5.5 Use case scenario: Analysis of 2016 United States presidential elections

In order to demonstrate how useful the proposed approach is, we introduce a possible use case scenario for the designed political discourse classifier: to analyse the tweets of the presidential (Hillary Clinton and Donald Trump) and vice-presidential (Tim Kaine and Mike Pence) candidates for the 2016 United States presidential elections.

We used a dataset of the 2016 United States Presidential Election Tweet IDs(Littman et al., 2016) with tweets gathered between July 13, 2016 and

Experiment	Accuracy	F-Measure
ML1	29.95% ±0.5	25.02 ±1.81
ML1 + Previous Tweet	30.71% ±1.32	24.5 ±1.6
ML1 + Political party (One hot)	30.85% ±0.6	24.4 ±1.5
ML1 + Political party (Disentangled)	30.17% ±1.03	25 ±1.48
ML1 + Political party (One hot) + Previous Tweet	**31.32**% ±1.27	25.86 ±1.2
ML1 + Political party (Disentangled) + Previous Tweet	31.18% ±1.22	**25.97** ±2.47
ML2	**39.6**% ±1.7	41.42 ±1.93
ML2 + Previous Tweet	38.74% ±2.22	40.7 ±4
ML2 + Political party (One hot)	39.35% ±	40.5 ±3.8
ML2 + Political party (Disentangled)	39.1% ±1.5	41.23 ±3.3
ML2 + Political party (One hot) + Previous Tweet	39.16% ±1.58	41.43±2.1
ML2 + Political party (Disentangled) + Previous Tweet	39.16% ±0.5	**42.34** ±1.4
ML3	37.35% ±2.58	39.25 ±3
ML3 + Previous Tweet	37.87% ±1.14	39.71 ±2
ML3 + Political party (One hot)	**39.27**% ±0.9	39.12 ±3.13
ML3 + Political party (Disentangled)	38.91% ±1.09	41.27 ±1.14
ML3 + Political party (One hot) + Previous Tweet	39.23% ±1.87	40.13 ±1.4
ML3 + Political party (Disentangled) + Previous Tweet	38.88% ±2.74	**42.71** ±2.14

Table 5.17: Multilabel subdomain results with BERT with a strict evaluation(the average results of 5 runs per experiment are shown) for each one of the experiment configuration. The accuracy (acc), F-Measure(macro) and their respective standard deviation is shown.

Experiment	Accuracy	F-Measure
ML1	42.1% ±1.1	25.41 ±1.64
ML1 + Previous Tweet	41.04% ±0.8	25.4 ±0.7
ML1 + Political party (One hot)	40.3% ±1.37	24.51 ±1.75
ML1 + Political party (Disentangled)	40.72% ±0.9	26.3 ±1.4
ML1 + Political party (One hot) + Previous Tweet	**43.78%** ±1.83	26.52 ±1.17
ML1 + Political party (Disentangled) + Previous Tweet	43.15% ±1.89	**27.58** ±2.56
ML2	55.37% ±0.9	43.72 ±1.33
ML2 + Previous Tweet	**56.2%**±1.4	**46.52** ±4.3
ML2 + Political party (One hot)	55.53% ±0.6	43.67 ±1.86
ML2 + Political party (Disentangled)	55.06% ±0.5	45.54±3.6
ML2 + Political party (One hot) + Previous Tweet	55.46% ±0.9	45.83 ±3.06
ML2 + Political party (Disentangled) + Previous Tweet	55.49% ±1.08	46.1 ±1.43
ML3	54.22% ±1.38	42.87 ±5.87
ML3 + Previous Tweet	**54.78%** ±1.07	43 ±2.1
ML3 + Political party (One hot)	55.15% ±1.02	45.25 ±2.8
ML3 + Political party (Disentangled)	55.37% ±1.42	46.32±1.38
ML3 + Political party (One hot) + Previous Tweet	55.04% ±2.58	44.52 ±3.18
ML3 + Political party (Disentangled) + Previous Tweet	54.75% ±2.1	**47.5** ±2.3

Table 5.18: Multilabel subdomain results with BERT with a less strict evaluation(the average results of 5 runs per experiment are shown) for each one of the experiment configuration. The accuracy (acc), F-Measure(macro) and their respective standard deviation is shown.

November 10, 2016. However, we only used a small part of the dataset: presidential and vice-presidentials candidates' timelines (ignoring RTs) during the previously mentioned time period: 5346 tweets from Hillary Clinton, 3364 from Tim Kaine, 4510 from Donald Trump and 1744 from Mike Pence. We processed candidates' tweets with the same procedure used for the annotated tweets: tokenization, removing stopwords and URLs and maintaining hashtags.

First of all, we performed a preliminary analysis classifying candidates' tweets in the previously mentioned 7 high level policy domains in order to have a general overview of each political parties (democratic and republican) preferences. To do so, we used a model trained with political manifestos and finetuned with annotated political tweets.

Furthermore, the political affiliation of the transmitter and the previous tweet was used as contextual data (the best results were achieved using the political leaning as an extra feature, see Table 5.14).

In Figure 5.2 the distribution of tweets of the Republican (red) and Democratic (blue) parties over the 7 high level policy domains can be seen.

The first worth mentioning aspect is how *Political System* is the dominant category for Republicans, whereas *Welfare and Quality of Life* is for Democrats. However, the Democratic party also emphasises in the Political system being their second priority. One of the reasons behind this could be that inside the high level *Political System* domain, there is a category named *Political Authority* which encompasses messages related with politician's competence to govern or the political opponent's lack of such competence. Therefore, tweets complimenting his or her allies and criticising his or her opponents would belong to *Political System* domain. Concerning the *Fabric of Society* domain, Republicans emphasise more than democrats in this high level policy. In the rest of high level policy domains, both parties have similar distributions.

However, this kind of political discourse analysis based on 7 high policy domains does not offer an accurate view of what is really happening, it only offers a general overview. This is the reason why we are proposing a more precise approach for political discourse analysis in Social Networks using the 56 categories defined in Table 3.3.

Figure 5.2: Distribution among 7 high level domains of the tweets created by Democratic (blue) and Republican (red) candidates.

Therefore, we have applied this new approach for analysing 2016 presidential elections. Nonetheless, in this use case we are going to emphasise on those categories that are not marginal. Marginal categories are those with less than 0.5% of the total amount of tweets in both political parties.

In Figure 5.3 what highlights the most is the fact that more than 25% of the tweets from both parties have been classified as *Political Authority* (305), which means that most of the political discourse during this elections was focused on attacking the opponent or praising themselves. However, the discourse from the republicans was more *Political Authority* centred compared to Democratic Party. This results coincide with the results obtained manually by (Russell, 2018).

Another point of interest would the disparity between republicans and democrats regarding *Equality* (503) category, which includes policies related with social justice, fair treatment of all people and the end of discrimination

according to the Manifestos' Project handbook[1]. It is also remarkable that this disparity can also be detected with *Welfare State Expansion* (504) category in a similar way. Moreover, republicans talk more about *Welfare State Limitation* (505) than democrats do.

In respect of their policy preferences regarding Economy, Republicans focused on tweets about *Free Market Economy (401)*, whereas Democrats where more concerned about *Market Regulation (403)*. Nonetheless, both parties have similar percentages of tweets related with *Incentives (402)* for businesses.

Regarding nationalism and immigration, it can be seen clearly in those categories related with immigration that Republicans sent more anti-immigration tweets than Democrats.

For instance, 5.26% of the Republican tweets have been classified in *the National Way Of Life - Positive* (601) category, where statements about patriotism and against the process of immigration are included, unlike Democrats whose 1.96% of the tweets are related to this matter. Moreover, Republicans have a marginal representation in those categories that promote immigrants' rights: *National Way of Live - Negative* (602) and *Multiculturalism* (607).

To conclude, even though there are small differences, Republicans were more concerned about *Military: Positive* (104) and *Law and Order* (605) categories than Democrats. However, the opposite happens with *Environmental protection (501) and Labour Groups - Positive* where democrats shown more interest (701).

[1] https://manifesto-project.wzb.eu/down/papers/handbook_2014_version_5.pdf

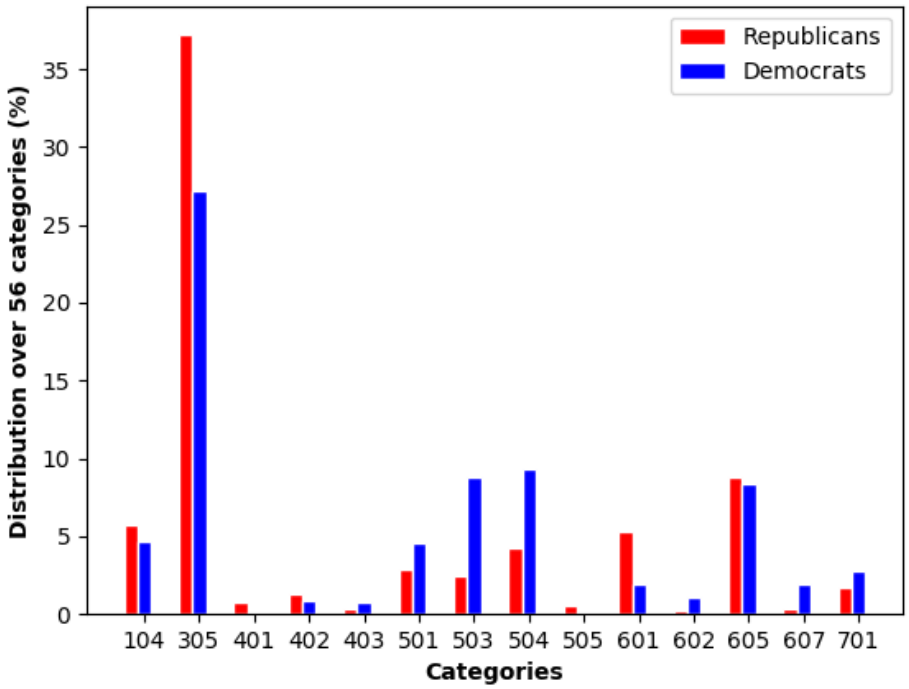

Figure 5.3: Distribution among 56 subdomains of the tweets created by Democratic (blue) and Republican (red) candidates.

The greatest teacher, failure is.

Yoda

CHAPTER

6

Conclusions and Future Work

A summary of the main results and contributions presented during this dissertation are detailed in this chapter. Therefore, to finalise this dissertation the objectives given in Chapter 1 are reviewed in order to evaluate at what level those objectives have been achieved. Moreover, a review of the main contributions made by this research work, along with a list of publications related to this PhD dissertation in order to show that this research work has been validated by the research community. The chapter ends with some ideas for future research in the area of political discourse analysis based on annotated political manifestos.

The rest of this chapter is structured as follows: Section 6.1 summarises the work done and conclusions obtained from this dissertation. Section 6.2 lists the main contributions of this dissertation. Section 6.3 explains how the objectives stated at the beginning of this dissertation has been achieved. Section 6.4 lists all scientific publications published during the development of this dissertation. Section 6.5 introduces possible future research. Section 6.6 makes some final remarks.

6.1 Summary of Work and Conclusions

To sum up, in this PhD dissertation we have taken a widely used content analysis technique designed by CMP for studying election manifestos and we have adapted it for its automated application in any type of political text with two main goals: to help political scientists in this complicated and time-consuming task, achieving state of the art results in automated manifestos classification; and applying this methodology on Twitter, analysing Tweets using the same categorisation scheme political scientists have used since 2002 to study election manifestos from all over the world. In both cases, we have used two types of contextual data, what has been said previously in the analysed fragment and who has said it, in order to enhance supervised classifiers performance.

The seed of this PhD dissertation was the identification of a dataset with a large number of annotated political manifestos that had not been used out of their research purpose: content analysis of election manifestos in order to study the policy preferences of each political parting taking solely into account their manifestos. Therefore, we identified a research area with plenty of work to do, since as it can be seen in Chapter 2, Related Work, there are not research works to the best of our knowledge, that have completed an in depth political discourse analysis as the one that can be performed applying CPM categorisation scheme in Social Networks.

Regarding the automated utilisation of the annotated election manifestos provided by CPM, when this PhD dissertation started there was an attempt of automated classification of election manifestos using its highest granularity, Domains. However, there were not applications of this large dataset outside manifestos. Nonetheless, as this dissertation has progressed and time has passed new research works have been developed: new attempts for automated classification of election manifestos using subdomains or even an application outside manifestos, the analysis of political speeches. The state of the art review in this field is available in Chapter 3.

However, as it has been previously clarified, our goal was to use a validated content analysis methodology and use it in Social Networks such as Twitter. To do so, we divided our research in two parts: first, help in the automation

of manifestos annotation process and second, use the acquired knowledge and apply it in social networks. In this stage, we identified two contextual information that we thought could be useful when it comes to classifying both manifestos and tweets: what has been said previously and by who. Chapter 4 explains more thoroughly the reasoning behind this decision and the different types of methods for parties' representation tested during this dissertation.

The evaluation of the proposed approach has been made in Chapter 5. Among our most important finding are that we have improved state of art results in 4 out of 7 languages for automated manifestos classification using our approach based on contextual information. Also, our approach is complementary to the other approaches used in the state of the art. Therefore, combining those approaches with our contribution, we could improve the results. Moreover, we have statistically validated that used contextual information does improve classifier performance. The validation has been performed using the Friedman test with the corresponding post-hoc tests (Nemenyi test). Furthermore, we have designed a new method for political parties' modelling in neural networks, using their political orientation in order to create a disentangled representation of each party. This new method has achieved the best results, outperforming other methods based on RILE score (left-right axis) or treating each of the political parties as unique.

Regarding the performance of our approach for tweets, the best results have once again being achieved using the previously presented contextual information: previous tweet and the representation of the political party. Moreover, we have been able to prove how annotated political manifestos and annotated political tweets are complementary information when it comes to training our political discourse classifier. Finally, in Section 5.5 , we have introduced a use case scenario explaining how our approach could be used to analyse the political discourse in Social Networks, analysing the 2016 United States presidential elections on Twitter.

6.2 Contributions

A summary of the contributions explained in this dissertation is presented in this section:

- *It has been statistically certified that adding as contextual information, the previous phrase improves the performance of the classifier when automatically annotating political manifestos.* This contribution addresses the objective 2, where the model able to have different type of inputs is designed and the objective 3, where the performance of this contextual information is analysed.

- *It has been statistically certified that adding as contextual information, the political party to which the phrase belongs, improves the performance of the classifier when automatically annotating political manifestos.* This contribution addresses the objective 2, where the model able to have different type of inputs is designed and the objective 3, where the performance of this contextual information is analysed.

- *State of the art results have been improved in 4 out of 7 languages when classifying subdomains in the automated manifestos annotation task.* This contribution addresses the objective 1 where the state of the art review has been performed and objective 2 where the used supervised classification models have been designed. Moreover, we have reached near human-level performance.

- *A new representation method for the use of political parties as input feature in supervised classifiers using a disentangled representation based on their political orientation. This representation allows the addition of new unknown parties to the model without any retraining effort.* This contribution addresses the objective 3.

- *It has been proven how annotated political manifestos and annotated political tweets are complementary information when it comes to training the political discourse classifier.* Addresses the objective 6 where an online political discourse classifier has to be designed.

- *It also has been proven that using the previous tweet and political party as additional contextual data achieves the best results classifying annotated tweets.* This contribution also refers to objective 6.

- *A novel approach for automatically classifying political tweets using a categorisation scheme widely used by political scientists.* With the findings achieved during objective 6, we have analyse the 2016 United States presidential elections using the best designed model.

- *An analysis of how would perform a multi-label political discourse classifier in order to measure its feasibility.* This contribution has been achieved following objectives 6 and 7.

- *A dataset of 5,000 tweets annotated with the CPM coding schema has been created.* In order to evaluate the performance of our political discourse classifier we annotated thousands of political tweets. This contribution has been achieved following objectives 5 and 6.

- Word2Vec embedding models for the Spanish language from text recovered from news, Wikipedia, the Spanish BOE, web crawling and open literary sources. This contribution has been achieved following objectives 2 and 6, in the design and building of the political discourse classifier.(Almeida and Bilbao, 2018)

6.3 Hypothesis and objective validation

In the beginning of this dissertation, concretely in Section 1.2, a hypothesis was posed, which stated the following:

> **Hypothesis 1** *Using contextual information it is possible to improve the automated election manifestos annotation process and perform a political discourse analysis in on-line social networks using manifestos' annotation scheme and the same contextual data previously used.*

In order to be able to validate this hypothesis, a goal was also defined, which is also shown below for convenience:

Goal 1 *To design and implement a political discourse classifier that uses annotated political manifestos, a very reduced amount of annotated political tweets and the context of each of those tweets to analyse on-line political discourse.*

1. *To study the current start of the art on political discourse analysis in social Networks and the automated used of annotated political manifestos.* In Chapter 2, the most relevant works about political discourse analysis in Social Networks have been reviewed. This analysis has been divided in two sections: manual and automated approaches. Then, in Chapter 3 the theoretical foundations of CMP methodology is described and the research works in the automated used of political manifestos are reviewed, explaining why our approach differs from them and how could contribute to the state of the art.

2. *To design and implement a supervised classification model for text categorization optimized for the problem and able to have different inputs than raw text.* In Section 4.2, the used text classification models are thoroughly explained: why CNNs and BERT are the best choices for this task, how has the contextual information been added as input to the models, etc.

3. *To identify an appropriate evaluation methodology for the automated manifestos annotation task with its corresponding metrics and perform a quantitative analysis of the results.* In Section 5.2 a global analysis of how this type of tasks should be evaluated is performed. We also criticise how automated manifestos annotation task have been evaluated so far. In 5.3.1, the followed experimental and evaluation setup of this specific task is explained.

4. *To analyse how the added contextual data affects supervised classifier's performance when classifying election manifestos.* In Section 5.3.3 we show how we have statistically validated the improvement given by the two proposed contextual data, achieving state of the art results in 4 out

7 languages. Moreover, we also validate our new representation method for parties using their political orientation.

5. *To identify an appropriate evaluation methodology for the on-line political discourse analysis task with its corresponding metrics and perform a quantitative analysis of the results.* Based on the global analysis made in Section 5.2, Section 5.3.1 explains how is the used dataset for the evaluation and how it has been performed: the experimental setup and analysed metrics.

6. *To analyse if annotated political manifestos could be used as complementary data to the annotated tweets in order to improve the performance of the political discourse classifier.* In Section 5.4, the results obtained after fine-tuning with annotated tweets a model already trained with annotated manifestos can be seen. The results achieved after the fine-tuning process are the best performing ones.

7. *To analyse how does the designed approach analyse the on-line political discourse using contextual information.* In Section 5.5 a use case scenario is introduced where an analysis of 2016 United States presidential elections is performed using the designed approach. The results achieved with approach are discussed in Section 5.4.3 where it has also been validated that the contextual information improves political discourse classifier's performance which obtains the best results after being fine-tuned with annotated tweets. Also, in Section 5.4.3 we have discussed about the feasibility of addressing this problem as a multi-label classification task.

Finally, through the accomplishment of the objectives and the goal, the hypothesis of this dissertation has been validated. The results obtained in Chapter 5 prove that it is possible to automate the election manifestos annotation process using contextual information and that is possible to use this knowledge to later perform a political discourse analysis in on-line social networks using manifestos' annotation scheme and the same contextual data.

6.4 Relevant Publications

During the development of this dissertation, the following scientific manuscripts have been presented to the scientific community and published in relevant international forums, such as indexed journals and conferences.

6.4.1 International JCR Journals

The analysis of how contextual information such as the previous phrase or the representation of the political party improves and achieves state of the art performance in automated manifestos classification task.

- Aritz Bilbao Jayo, Aitor Almeida. (2018) "Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data" In International Journal of Distributed Sensor Networks. DOI: 10.1177/1550147718811827. JCR Impact Factor (2018): 1.614 , Q3. November 2018.

The preliminary work that led to this dissertation was published in the following journal:

- Aritz Bilbao Jayo, Aitor Almeida, Diego López-de-Ipiña. (2016) "Promotion of active ageing combining sensor and social network data" In Journal of Biomedical Informatics. vol. 64. p. 108-115. DOI: 10.1016/j.jbi.2016.09.017. JCR Impact Factor (2016): 2.447 , Q1. October 2016.

Also, a parallel work focused in the Smart Cities and the potential application of this work in the field was published:

- Ruben Sánchez Corcuera, Adrian Núñez-Marcos, Jesus Sesma-Solance, Aritz Bilbao Jayo, Rubén Mulero, Unai Zulaika Zurimendi, Gorka Azkune, Aitor Almeida. (2019) "Smart cities survey: Technologies, application domains and challenges for the cities of the future" In International Journal of Distributed Sensor Networks. vol. 15. p. 1550147719853984. DOI: 10.1177/1550147719853984. CR Impact Factor (2018): 1.614 , Q3. June 2019.

6.4.2 International Conferences

A first approach for analysing political discourse in social networks was presented.

- Aritz Bilbao Jayo, Aitor Almeida. (2018) "Political discourse classification in social networks using context sensitive convolutional neural networks" In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media. p. 76-85.

As a consequence of the work performed during this PhD Dissertation, NLP techniques were adapted for behaviour modelling achieving promising results.

- Aitor Almeida, Gorka Azkune, Aritz Bilbao Jayo. (2018) "Embedding-level attention and multi-scale convolutional neural networks for behaviour modelling" In Proceedings of conference: Ubiquitous Intelligence and Computing 2018 (UIC 2018). DOI: 10.1109/SmartWorld.2018.00103.

6.4.3 Datasets

- Dataset of 5,000 tweets annotated with the CPM coding schema.[1]

6.4.4 Technical Contributions

- Word2Vec embedding models for the Spanish language(Almeida and Bilbao, 2018).
- The source code of this Dissertation will be released in Github. [2]

6.5 Future work

Inspired by the limitations of the research presented in this dissertation, we have identified the following further research lines:

[1] https://github.com/AritzBi/tweets_manifestos_methodology_2016_usa
[2] https://github.com/AritzBi/manifestos_context_classifier

- In this research work, we have focused in the main political topic or idea a tweet could contain, a multi-class classification problem. However, we have also analysed the results of a multi-label classification problem, assuming that a tweet could contain more than one idea. However, the results obtained with this assumption are considerably worse than the metrics obtained considering that each tweet represent a principal political idea. Therefore, we believe that in order to improve the results in the multi-label classification task, first, a bigger dataset of annotated tweets should be needed, and second, a multi-label specific neural network architectures should be tested.

- Even though during this dissertation the main objective was to bring to Social Networks an election manifestos content analysis methodology such as the CMP, we have improved state of the art results in 4 out of 7 languages in the automated annotation of election manifestos. The best results obtained in the remaining 3 languages were achieved using a cross-lingual classification approach, where languages with more annotated manifestos enhanced those languages with less training samples. Therefore, we expect that combining the previously defined approach with the contextual information introduced during this PhD dissertation, better results would be achieved. Moreover, following this approach, the disentangled representation of the parties could be enriched since the number of parties' representation training together would increase (in this Dissertation parties are divided by language) and therefore, it could increase this contextual information's performance.

- It would be interesting to analyse the subdomain *305 Political Authority* in Social Networks from a positive or negative point of view. This category encompasses those statements with a partisan rhetoric where politicians praise their policies or actions, whereas criticise their rivals. Unfortunately, this category does not differentiate the first from the latter as (Russell, 2018) did. Moreover, it would interesting to apply Named Entity Recognition (NER) techniques in this category in order to analyse who are they talking about and how. Thus, we consider

that this addition would enrich the political discourse analysis in Social Networks.

6.6 Final Remarks

With this dissertation we have tried to adapt a validated and widely used content analysis methodology for election manifestos into a new area, the Social Networks, where politicians spread their ideas more frequently than in election manifestos. As a result of this new context, the amount of written statements from politicians have augmented considerably in the last years. Therefore, the traditional approach used by political scientists of manually annotating political texts is not feasible in order to analyse all the political written statements generated every day. Thus, we strongly believe that the future of this field of research lies on a joint effort between political scientists and natural language processing researchers.

Bibliography

Almeida, A. and Bilbao, A. (2018). Spanish 3b words word2vec embeddings. `https://doi.org/10.5281/zenodo.1155474`. 9, 44, 99, 103

Alonso, S., Cabeza, L., and Gómez, B. (2017). Disentangling peripheral parties' issue packages in subnational elections. *Comparative European Politics*, 15(2):240–263. 1, 25

Alonso, S., Gómez, B., and Cabeza, L. (2013). Measuring centre–periphery preferences: The regional manifestos project. *Regional & Federal Studies*, 23(2):189–211. 27

Barandela, R., Sánchez, J. S., Garca, V., and Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851. 61

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828. 3, 40

Benoit, K. (2009). Irish political parties and policy stances on european integration. *Irish Political Studies*, 24(4):447–466. 1, 24

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., and Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110(2):278–295. v, 35, 36

Benoit, K. and Laver, M. (2007). Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies*, 26(1):90–107. 25

Bilbao-Jayo, A. and Almeida, A. (2018). Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, 14(11):1550147718811827. 76

Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118. 48

Brunsbach, S., John, S., and Werner, A. (2012). The supply side of second-order elections: Comparing german national and european election manifestos. *German Politics*, 21(1):91–115. 25

Budge, I. (2001). *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998*, volume 1. Oxford University Press on Demand. 1, 24

Budge, I. (2013). The standard right-left scale. *London: Essex University*. 1

Budge, I. and Laver, M. J. (2016). *Party policy and government coalitions*. Springer. 25

Casero-Ripollés, A., Sintes-Olivella, M., and Franch, P. (2017). The populist political communication style in action: Podemos's issues and functions on twitter during the 2016 spanish general election. *American behavioral scientist*, 61(9):986–1001. 15, 16

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. *ICWSM*, 133:89–96. 17

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30. 63

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. v, 38, 50, 52, 53

Finn, S., Mustafaraj, E., and Metaxas, P. T. (2014). The co-retweeted network and its applications for measuring the perceived political polarization. 17

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92. 63

Gayo Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence. 17

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017). Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 42–46. 29, 30, 34

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. 48

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*. 48

Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political ideology detection using recursive neural networks. In *Proceedings of the Association for Computational Linguistics*, pages 1113–1122. 18

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*. 38, 42, 45

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 49

LeCun, Y. et al. (2015). Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, page 20. 45

Lehmann, P., Matthieß, T., Merz, N., Regel, S., and Werner, A. (2018). Manifesto corpus. version: 2017-2. berlin: Wzb berlin social science center. vii, 62, 63

Littman, J., Wrubel, L., and Kerchner, D. (2016). 2016 United States Presidential Election Tweet Ids. 88

López-García, G. (2016). 'new'vs' old'leaderships: the campaign of spanish general elections 2015 on twitter. *Communication & society*, 29(3). 14, 16

Lowe, W., Benoit, K., Mikhaylov, S., and Laver, M. (2011). Scaling policy preferences from coded political texts. *Legislative studies quarterly*, 36(1):123–155. 1

Mikhaylov, S., Laver, M., and Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91. 2, 35, 75

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 44

Nanni, F., Zirn, C., Glavas, G., Eichorst, J., and Ponzetto, S. P. (2016). Topfish: Topic-based analysis of political position in us electoral campaigns. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016)*. 29, 34

Nemenyi, P. (1962). Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210. 65

Nordsieck, W. (2015). Parties and elections in europe. *Parties and Elections in Europe*. 40

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2. 16

Park, K. (2018). Pre-trained word vectors of 30+ languages. `https://github.com/Kyubyong/wordvectors`. 44

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*. 50

Poria, S., Cambria, E., and Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544. 42

Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49. 42

Prechelt, L. (1998). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer. 62, 79

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI. 50

Ramos-Serrano, M., Fernandez Gomez, J. D., and Pineda, A. (2018). 'follow the closing of the campaign on streaming': The use of twitter by spanish political parties during the 2014 european elections. *New media & society*, 20(1):122–140. 14, 16

Rao, A. and Spasojevic, N. (2016). Actionable and political text classification using word embeddings and lstm. *arXiv preprint arXiv:1607.02501*. 18

Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Detecting and tracking political abuse in social media. *ICWSM*, 11:297–304. 18

Russell, A. (2018). Us senators on twitter: Asymmetric party rhetoric in 140 characters. *American Politics Research*, 46(4):695–723. 15, 92, 104

Schmitt-Beck, R., Bytzek, E., Rattinger, H., Roßteutscher, S., and Weßels, B. (2009). The german longitudinal election study (gles). *Vortrag im Rahmen der Jahrestagung der International Communication Association (ICA), Chicago*, 21:25. 19

Scikit-Learn-Developers. *Scikit Learn User Guide 2019*. 61

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437. 60

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958. 49

Stier, S., Bleier, A., Lietz, H., and Strohmaier, M. (2018). Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political communication*, 35(1):50–74. 19, 21

Subramanian, S., Cohn, T., and Baldwin, T. (2018). Hierarchical structured model for fine-to-coarse manifesto text analysis. *arXiv preprint arXiv:1805.02823*. 30, 34, 74, 76

Subramanian, S., Cohn, T., Baldwin, T., and Brooke, J. (2017). Joint sentence-document model for manifesto text analysis. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 25–33. 30, 34, 76

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. 16

Valverde-Albacete, F. J., Carrillo-de Albornoz, J., and Peláez-Moreno, C. (2013). A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 41–52. Springer. 58

Van Asch, V. (2013). Macro-and micro-averaged evaluation measures. 60

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. v, 50, 54

Volkens, A. (2002). Manifesto coding instructions. `https://www.poltext.org/sites/poltext.org/files/iii02-201.pdf`. 24

Volkens, A., Krause, W., Lehmann, P., MatthieÃŸ, T., Merz, N., Regel, S., and WeÃŸels, B. (2019). The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2019a. vii, 2, 24, 32, 33

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. 51

Wüst, A. M. and Volkens, A. (2003). Euromanifesto coding instructions. `http://www.mzes.uni-mannheim.de/publications/wp/wp-64.pdf`. 1

Yaqub, U., Chun, S. A., Atluri, V., and Vaidya, J. (2017). Analysis of political discourse on twitter in the context of the 2016 us presidential elections. *Government Information Quarterly*, 34(4):613–626. 20

Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*. 48, 49

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27. 50

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. `https://ub-madoc.bib.uni-mannheim.de/41552/1/classyman.pdf`. 28, 34

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., and Lukasik, M. (2016a). Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028*. 19

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., and Tolmie, P. (2016b). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989. 19

APPENDIX

A

Parties and their Political Orientations

Party	Orientation	Party	Orientation
Liberal Alliance	Liberalism	Red-Green Unity List	Socialism Eco-Socialism Euroscepticism
Socialist People's Party	Green Politics Democratic socialism	Social Democratic Party	Social Democracy
Centre democrats	Liberal Conservatism Centrism	Danish Social-Liberal Party	Social liberalism
Liberals	Conservative liberalism Agriarinism	Christian People's Party	Christian Democracy
Conservative People's Party	Liberal Conservatism	Danish People's Party	National conservatism Social Convervatism RIght-wing populism
Progress Party	RIght-wing populism Economic liberalism	Green Union	Green Politics
Left Wing Alliance	Eco-Socialism Democratic socialism	Finish Social Democrats	Social Democracy
Finnish Christian Union	Christian Democracy Social Convervatism	National Coalition	Liberal Conservatism
Centre Party	Agriarinism Centrism	Finnish Rural Party	Conservatism Social Convervatism RIght-wing populism
Left Front	Socialism Communism	Europe Ecology The Greens	Green Politics

French Communist Party	Communism	Left Radical Party	Social liberalism
Indomitable France	Euroscepticism Alter-Globalisation Left Wing Populism Democratic socialism	Socialist Party	Social Democracy Democratic socialism
Radical party	Liberalism	Republic Onwards	Liberalism Social liberalism Pro Europeanism
Union of Democrats and Independents	Liberalism	Democratic Movements	Centrism
The Republicans	Liberal Conservatism Gaullism	New Centre	Conservative liberalism Centrism
Centrist Alliance	Liberalism Pro Europeanism Centrism	National Front	RIght-wing populism Nationalism
Geneva Citizens' Movement	RIght-wing populism Regionalism	Civil Revolution	Euroscepticism Green Politics Communism Anti-Corruption Populism
People of Freedom	Liberal Conservatism Christian Democracy	Left Ecology Freedom	Eco-Socialism Democratic socialism
Democratic Party	Social Democracy Christian Left	Democratic Centre	Centrism
Civic Choice	Liberalism	Union of the Center	Christian Democracy Social Convervatism
Brothers of Italy	National conservatism Nationalism	Labour and Freedom List	Euroscepticism Christian Democracy National conservatism RIght-wing populism Isoliatiosnm
Northern League	RIght-wing populism Regionalism	List Di Pietro Italy of Values	Centrism Anti-Corruption
Autonomy Progress Federalism Aosta Valley	Regionalism Autonomism Big Tent	Five Star Movement	Christian Democracy Centrism Anti-Corruption Direct Democracy Enviromentalism
Ticino League	Euroscepticism National conservatism RIght-wing populism Regionalism Isoliatiosnm	South Tyrolean People's Party	Christian Democracy Minority Interests
Alliance'90	Green Politics	Party of Democratic Socialism	Democratic socialism
The Left. Party of Democratic Socialism	Democratic socialism	The Left	Democratic socialism
Social Democratic Party of Germany	Social Democracy	Free Democratic Party	Liberalism
Christian Democratic Union	Liberal Conservatism Christian Democracy	Pirates	Direct Democracy Copyright reform Transparency
Alternative for Germany	National conservatism	The Greens	Green Politics

Austrian Communist Party	Socialism Comunism	Austrian Social Democratic Party	Social Democracy
Austrian Freedom Party	National conservatism RIght-wing populism	The New Austria and Liberal Forum	Liberalism
Austrian People's Party	Conservatism Christian Democracy	Alliance for the Future of Austria	RIght-wing populism Economic liberalism
Team Stronach for Austria	Euroscepticism Economic liberalism	Green Party of Switzerland	Green Politics
Green Liberal Party	Liberalism Enviromentalism	Swiss Labour Party	Socialism Communism
Social Democratic Party of Switzerland	Social Democracy	Radical Democratic Party	Liberalism
Christian Democratic People's Party of Switzerland	Christian Democracy	Protestant People's Party of Switzerland	Christian Democracy Social Convervatism Evangelicalism
Christian Social Party	Christian Democracy Regionalism	Federal Democratic Union	National conservatism Social Convervatism Christian right
Swiss People's Party	National conservatism RIght-wing populism Economic liberalism	Conservative Democratic Party of Switzerland	Conservatism
Economic Freedom Fighters (South Africa)	Communism Left Wing Populism Marxism-Leninism Comunism Anti-capitalism Anti-imperialism Pan-Africanism, Anti-Europeanism	African National Congress	Social Democracy Anti-imperalism african nationalism
Democratic Party	Liberalism Conservative liberalism Anti-apartheid	Democratic Alliance	Liberalism Social liberalism Centrism
Congress of the People	Social liberalism Social Democracy progressivism	Inkatha Freedom Party	Conservatism Economic liberalism Federalism Anti-communism Zulu nationalism
Green Party of England and Wales	Green Politics	We Ourselves	Democratic socialism United Ireland
Labour Party	Social Democracy	Social Democratic and Labour Party	Social Democracy United Ireland
Liberal Democrats	Liberalism Social liberalism	Conservative Party	Euroscepticism Conservatism Economic liberalism
Ulster Unionist Party	Conservatism Unionism	The Party of Wales	Social Democracy Separatism
Scottish National Party	Social Democracy Unionism	Democratic Unionist Party	National conservatism Social Convervatism Unionism
United Kingdom Independence Party	Euroscepticism RIght-wing populism Economic liberalism	United Left Alliance (Ireland)	Euroscepticism Democratic socialism
Green Party (Ireland)	Green Politics	Socialist Party (Ireland)	Trotskyism

People Before Profit Alliance (Ireland)	Socialism Trotskyism	Anti-Austerity Alliance (Ireland)	Socialism Trotskyism
Workers and Unemployment Action (Ireland)	Socialism Local Politics	Labour Party (Ireland)	Social Democracy
Social Democrats (Ireland)	Social Democracy	Progressive Democrats (Ireland)	Conservative liberalism Classical Liberalism
Familiy of the Irish	Liberal Conservatism Christian Democracy	Soldiers of Destiny	Conservatism
We Ourselves	Democratic socialism United Ireland	Independent Alliance	Big Tent Nonpartisan politics
Democratic Party (US)	Social liberalism Modern Liberalism	Republican Party (US)	Conservatism Social Convervatism
Australian Greens	Green Politics	Australian Labor Party	Social Democracy Democratic socialism
Palmer United Party	RIght-wing populism Australian Natinalism	Liberal Party of Australia	Liberalism Liberal Conservatism Social Convervatism Economic liberalism
Liberal National Party of Queensland	Liberal Conservatism Economic liberalism	Country Liberal Party	Liberal Conservatism Conservative liberalism Agriarinism
Katter's Australian Party	Conservatism Agriarinism Christian Democracy Social Convervatism RIght-wing populism Protecionism Australian Natinalism Rural Interests Economic Nationalism	Country Party	Conservatism Agriarinism
Green Party of Aotearoa New Zealand	Green Politics	New Zealand Labour Party	Social Democracy
ACT New Zealand	Classical Liberalism right-libertarianism	United Future New Zealand	Liberal Conservatism Centrism
Progressive Party	Democratic socialism progressivism	New Zealand National Party	Liberalism Liberal Conservatism Conservatism Economic liberalism
New Zealand First Party	Social Convervatism Centrism Nationalism Populism Protecionism	Māori Party	Maori Rights
Mana Party	Maori Rights tino-rangatiratanga	Social Credit Political League	Social Credit,
Popular Unity	Socialism Comunism	United We Can	Anti-Globalisation Democratic socialism Direct Democracy
Future Yes (Geroa Bai)	Separatism	Amaiur	Socialism Separatism
Commitment-Q	Socialism Green Politics Regionalism	Basque Country Unite	Socialism Separatism

We Can	Anti-Globalisation Democratic socialism Direct Democracy	United Left	Socialism Comunism
Spanish Socialist Workers' Party	Social Democracy	Citizens	Liberalism
Union, Progress and Democracy	Social liberalism	People's Party	Conservatism Christian Democracy
Convergence and Union	Liberalism Separatism	Forum Asturias	Conservatism Regionalism
Basque Nationalist Party	Christian Democracy Separatism	Basque Solidarity	Social Democracy Separatism
Andalusian Party	Social Democracy Andalusian Nationalism	Canarian Coalition and Canarian Nationalist Party	Centrism Regionalism
Aragonist Council	Socialism Eco-Socialism Regionalism	Navarrese People's Union	Conservatism Christian Democracy Regionalism

Table A.2: List of analysed parties with their respective political orientations.

APPENDIX

B

Examples of annotated manifestos

Manifesto	Sentence	Category
PSOE 2015	Definir un nuevo modelo de financiación de todas las enseñanzas profesionalizadoras que permita atender la demanda en las condiciones de calidad exigibles	411
PSOE 2015	El sistema universitario público ha mantenido durante décadas un ritmo de crecimiento tanto en alumnado como en profesorado y en titulaciones muy superior al resto de los países de nuestro entorno.	506
PSOE 2015	que solo será posible invocando la combinación efectiva de eficiencia con equidad, crecimiento verde, inclusión y protagonismo de la toda la sociedad	416.2
PSOE 2015	Asimismo, cada vez urge más articular medidas para actualizar y garantizar derechos propios de las sociedades democróticas, abordando fenómenos como la suplantación de la identidad, la privacidad o el ciberacoso	201.2
PSOE 2015	Impulsar el plan de Educación Digital para promover la utilización de contenidos, recursos y herramientas digitales en todos los niveles del sistema educativo	506
PSOE 2015	y garantizar la igualdad de oportunidades durante la etapa de aprendizaje	503
PSOE 2015	Es aquello que nos define, configura nuestro imaginario colectivo y se convierte en factor de cohesión social	502
PSOE 2015	Todo ello ha supuesto un empobrecimiento del sector cultural, que ha resistido gracias a la dedicación de profesionales y artistas	408
PSOE 2015	España tiene en la diversidad una seña de identidad valiosa que debe ser preservada y cuidada como un valor común	301

Table B.1: Examples of annotated sentences from manifestos in Spanish

Manifesto	Sentence	Category
Conservative Party 2015	And failing to control our debt would be more than an economic failing; it would be a moral failing, leaving our children and grandchildren with debts that they could never hope to repay	414
Conservative Party 2015	Industrial action in these essential services would require the support of at least 40 per cent of all those entitled to take part in strike ballots	702
Conservative Party 2015	Last year alone, 140,000 disabled people found work.	503
Conservative Party 2015	We will allow farmers to smooth their profits for tax purposes over five years, up from the current two years, to counter income volatility	703
Conservative Party 2015	We will back the institution of marriage in our society, enabling married couples to transfer £1,060 of their tax-free income to their husband or wife, where the highest earner is a basic rate taxpayer	603
Conservative Party 2015	We will abolish long-term youth unemployment, and make sure that all young people are either earning or learning	404
Conservative Party 2015	But it is not fair , on taxpayers, or on young people themselves, that 18-21 year-olds with no work experience should slip straight into a life on benefits without first contributing to their community	505
Conservative Party 2015	We know how important it is to preserve vital community assets such as pubs, town halls and sports facilities, so we will strengthen the Community Right to Bid that we created	606
Conservative Party 2015	We will always be a party that is open, outward-looking and welcoming to people from all around the world	607

Table B.2: Examples of annotated sentences from manifestos in English

Manifesto	Sentence	Category
Five Star Movement	Divieto per i parlamentari di esercitare un'altra professione durante il mandato	304
Five Star Movement	Vietare la nomina di persone condannate in via definitiva (es. Scaroni all'Eni) come amministratori in aziende aventi come azionista lo Stato o quotate in Borsa	605
Five Star Movement	Piano di mobilità per i disabili obbligatorio a livello comunale	705
Five Star Movement	Incentivazione dei mercati locali con produzioni provenienti dal territorio	416
Five Star Movement	Proibizione di costruzione di nuovi parcheggi nelle aree urbane	501
Five Star Movement	Il Parlamento non rappresenta più i cittadini che non possono scegliere il candidato, ma solo il simbolo del partito.	202
Five Star Movement	Abolizione del Lodo Alfano	304
Five Star Movement	Assume un ruolo di consumatore e di elettore passivo, escluso dalle scelte che lo riguardano	202
Five Star Movement	Abolizione dell'Ordine dei giornalisti	403

Table B.3: Examples of annotated sentences from manifestos in Italian

Manifesto	Sentence	Category
Christian Democratic Union	Wir wollen das bestehende Gute sichern und gemeinsam noch Besseres schaffen	601.1
Christian Democratic Union	Die Einführung des gesetzlichen Mindestlohns in Deutschland hat sich grundsätzlich bewährt	412
Christian Democratic Union	Wir wollen, dass Arbeitnehmer am Erfolg ihres Unternehmens besser teilhaben können	701
Christian Democratic Union	Wir wollen mehr Frauen in Führungspositionen in Wirtschaft und Verwaltung.	503
Christian Democratic Union	Das spart Zeit und Geld und ermöglicht zusätzliche Wertschöpfung	303
Christian Democratic Union	Autoritäre Staatssysteme sind auf dem Vormarsch, scheinbar stabile Staaten sind zerbrochen	202.1
Christian Democratic Union	Nach Airbus und Ariane wäre es ein weiteres großes europäisches Projekt	108
Christian Democratic Union	Das Existenzrecht und die Sicherheit Israels sind Teil der deutschen Staatsräson.	101
Christian Democratic Union	Die große Mehrheit ebenso wie ethnische und gesellschaftliche Minderheiten.	607.2

Table B.4: Examples of annotated sentences from manifestos in German

Manifesto	Sentence	Category
Social Democratic Party	Danmark fører en aktiv udenrigspolitik, og det skal vi blive ved med	107
Social Democratic Party	For det betyder, at det er sværere at lære dansk, og man er afhængig af offentlige ydelser	608.2
Social Democratic Party	Og vi har afskaffet brugerbetalingen til behandling af barnløshed	504
Social Democratic Party	Og vi har et ansvar for, at vores børn er godt rustet til fremtiden	706
Social Democratic Party	Vi bygger f.eks. en ny motorvej fra Holstebro til Herning, en ny storstrømsbro og en ekstra etape af Kalundborg-motorvejen	411
Social Democratic Party	Det er en klar forskel i dansk politik.	305.1
Social Democratic Party	Derfor er Danmark også et af verdens førende lande, når det handler om grønne teknologier.	416.2
Social Democratic Party	Kort sagt: Vi har lagt krisen bag os, og Danmark er tilbage på sporet	409
Social Democratic Party	Et land med solidaritet mellem mennesker	606.1

Table B.5: Examples of annotated sentences from manifestos in Danish

Manifesto	Sentence	Category
Left Radical party	La légalisation de l'usage du cannabis, sous contrôle de l'Etat	605.2
Left Radical party	Confiance des entrepreneurs et des investisseurs dans le soutien qui leur sera apporté par les pouvoirs publics	408
Left Radical party	La construction de logements doit être respectueuse des objectifs d'économie d'énergie.	501
Left Radical party	L'exécutif est en charge de leur application,le pouvoir judiciaire veille à leur respect et sanctionne les manquements qui y sont apportés.	202.1
Left Radical party	La facilitationdes allers et retours entre vie professionnelle et vie politique devraêtre l'un des objets du statut de l'élu que je souhaite.	701
Left Radical party	La crise de l'Europe atteint les citoyens, dont le scepticisme à l'égard de la construction européennes'est accru, quand ils n'y sont pas – pour une minorité d'entre eux – franchement hostiles.	108
Left Radical party	Je suis en faveur d'un nouveau traité européen qui viendrait refonder la gouvernance de l'Union,la rendre à la fois plus démocratique et plus efficace.	204
Left Radical party	Je m'opposerai àtoute mesure qui tendrait à substituer aux subventions les seules formes d'aide à caractère de prêts.	403
Left Radical party	La démultiplication des échanges scolaires internationaux	506

Table B.6: Examples of annotated sentences from manifestos in French

Manifesto	Sentence	Category
Swedish People's Party	och förkorta utbetalningsperioden i motsvarande grad	701
Swedish People's Party	Livsrytmen måste ge utrymme för frivilligverksamhet.	606.1
Swedish People's Party	Jämlikheten är ett fundament i tryggandet av rättsstaten där var och en ska behandlas på lika och rättvisa grunder.	202.1
Swedish People's Party	satsa på förebyggande vård och stödformer inom öppenvården för familjer med barn för att minimera omhändertaganden	603
Swedish People's Party	Finland ska vara en aktiv och ansvarsfull internationell aktör i utvecklings- och handelspolitiken	107
Swedish People's Party	men behöver inte alltid själv vara producenten	505
Swedish People's Party	jobba för att Finland och EU i sin utvecklingspolitik stödjer länders utveckling av en god och öppen förvaltning som motverkar korruption	304
Swedish People's Party	Vi vill slå vakt om landsbygdens potential för sysselsättning och företagande	402
Swedish People's Party	locka utländska studerande som avlagt examen i Finland att stanna kvar och jobba genom att slopa sexmånadersregeln för uppehållstillstånd efter avlagd examen	607.2

Table B.7: Examples of annotated sentences from manifestos in Finnish

Account	Sentence	Category
ACunninghamMP	It's okay for us staying warm indoors - I think we often forget the dedication of the public sector workers who are out there trying to keep the roads safe public sector workers who've seen the value of their pay drop considerably in recent years. Thank you	704
AJonesMP	Catching a bus helps to lower our emissions and helps reduce congestion on our streets. New electric buses from @harrogatebus will soon further that effect with zero emissions. That is why I am supporting Catch the Bus Week.	501
alancampbellmp	Universal message from folk of North Shields at The Net open day to our friends and family in Scotland please vote No!#bettertogether	601.1
ALewerMBE	Today I highlighted the important work that @NorthamptonBC @NptonHopeCentre do to combat street homelessness at the parliamentary debate https://t.co/YvTIe1QsIw	606.1
AlexChalkChelt	It's time for water companies, who pay out vast dividends and salaries to senior management, to act more in the public interest - or face the consequences. This speech by @michaelgove is spot on	403
Alison_McGovern	Two weeks today we will know if the Tories in general and the Chancellor specifically care about British kids. Stop the #universalcredit chaos or watch #childpoverty rise. The choice is yours @PhilipHammondUK	504/305
CatMcKinnell	Looking forward to meeting with @jeremycorbyn and @UKLabour parliamentary colleagues today to discuss putting words into action on tackling #antisemitism in @UKLabour https://t.co/ql0cbbmRIZ	503/305
CharlieElphicke	Welcome interest rate cut. But we also need to look at cutting taxes -especially for small businesses and basic rate taxpayers.	402
timkaine	Annual Equality VA dinner reinvigorates our commitment to advancing equality for LGBT Virginians. Glad to see so many friendly faces tonight	503
RepMichaelWaltz	Proud to vote for legislation elevating the Special Envoy to Monitor Combat #AntiSemitism to the rank of Ambassador - allowing them to report directly to @StateDept @SecPompeo as the primary advisor coordinator for U.S. efforts to combat anti-Semitism abroad. #StandWithIsrael	101/605.1
RepTorresSmall	@realDonaldTrump It is a tremendous problem we need to secure our border to prevent violent criminals traffickers from crossing it. This is why we should invest $49 million for 275 new Domestic Investigation agents to look into human trafficking drug	605.1
SenShelby	I'm honored to be named a "Guardian of Small Business" from @NFIB. Small businesses are the lifeblood of our economy.	402
RepTipton	We are delivering on the promises we made to the American people with a booming economy safer communities and a strengthened military. Learn how we are #betteroffnow by visiting the following link: https://t.co/5AXHo1kbCO https://t.co/uq1oltlnYw	305/104/605.1
SenKamalaHarris	Immigrants from countries across the globe - including and especially those from Haiti and all parts of Africa have helped build this country. They should be welcomed and celebrated not demeaned and insulted.	602.2

Table B.8: Examples of annotated tweets

APPENDIX

C

Open Sources

- Source code: `https://github.com/AritzBi/manifestos_context_classifier`
- Annotated tweets: `https://github.com/AritzBi/tweets_manifestos_methodology_2016_usa`

Declaration

I, Aritz Bilbao Jayo, herewith declare that this dissertation is my own original work, carried out as a doctoral student at the University of Deusto. All assistance received and notions from other sources have been identified as such, acknowledging their correspondent contributions and citing them properly.

This work contains no material which has been presented in identical or similar form to any examination board, except where due acknowledgement is made in the dissertation.

This dissertation was finished writing on February 28$^{\text{th}}$, 2020.