Automatic Detection of Patronizing and Condescending Language Towards Vulnerable Communities

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Carla Pérez-Almendros

May 2023

Cardiff University
School of Computer Science & Informatics

To my parents, who were the seed.

To my husband, who was the wings and the nest.

To my son, who was the horizon.

Abstract

This thesis is focused on the study and analysis of Patronizing and Condescending Language (PCL) towards vulnerable communities. Someone using PCL displays a superior attitude towards others, raising a feeling of compassion and pity. PCL feeds stereotypes, causes discrimination and reinforces inequalities.

In this work, we analyze how NLP can help us to detect and categorize PCL, while enhancing human understanding of such language. To achieve this, we introduce a novel task to the NLP community, namely the Detection and Categorization of PCL towards vulnerable communities. This thesis contributes valuable insights by providing annotated data, baselines, and qualitative analysis from various experiments. The work developed in this thesis started with the creation of the Don't Patronize Me! (DPM!) dataset, with paragraphs extracted from media sources. Each paragraph was annotated to identify PCL and the specific techniques employed to express the condescension. A taxonomy of PCL categories was also introduced to classify these techniques. We analyzed the effectiveness of language models in detecting and categorizing PCL, showing that non-trivial results can be achieved, but room for improvement remains. We furthermore explored the impact of prior knowledge through transfer learning, revealing that exposure to certain types of data can benefit PCL detection models. Additionally, we share insights gained from organizing a SemEval task focused on PCL detection, which demonstrated that a judicious combination of standard models and SoTA techniques can achieve remarkable results. However, a closer look at the dataset unveiled that there are two types of PCL, namely linguistic

and thematic, and that the training data significantly influences the model's ability to detect specific PCL types. Overall, our findings confirm that language models can detect and categorize PCL to some extent, but specific approaches tailored to its unique characteristics are necessary. These findings improve our understanding of PCL and offer directions for future research.

Acknowledgements

This PhD journey has been one of the most exciting, challenging and rewarding experiences of my life. Four years ago, I decided to leave my job, my comfort zone and my previous career to start a PhD in a new discipline, in a second language and in a foreign country. As if these were not enough challenges, a global pandemic and a newborn child joined along the journey to make the adventure even more exciting. Today, I am writing the last lines of my thesis in NLP, and this would not have been possible without the support of the people who have accompanied me during all this experience, helping me to overcome the difficulties and celebrating the triumphs with me.

First of all, I need to thank my supervisor, Professor Steven Schockaert, for his priceless support. Thanks for believing in me from the beginning, even when I did not. Thanks for the many hours of explanations, discussions and thought-provoking conversations, for your patience and your enthusiasm. I know now how lucky I have been to have you as my supervisor. Thanks.

To all the staff and colleagues in the Cardiff NLP group, thanks for your help and words of support. It has been a pleasure to share this adventure with you all. I would also like to thank the School of Computer Science and Informatics, the ARCCA team and the Kaggle Open Data Research Grant for making my research possible.

I would like to thank now my husband and words get too small to tell about all I have to thank him. Thank you, Luis, for being my partner in life and in all its adventures,

for being always there, for helping me overcome my failures and frustrations and for making me acknowledge my achievements. Thanks for making us, together, the best team. And to our son, Luis Jr., who learnt to say "thesis" earlier than he should have, thanks for filling every moment (even the sleepless nights) with joy.

I want to extend my thanks to my family, who has been the seed of everything. Thank you for your support, your interest and your understanding. I could not have made it without knowing that you were by my side. To my friends, I want to say that maybe you did not know, but your words of support meant much more than you could imagine. They gave me strength and motivation every time.

To conclude, I just have a message for the scared, insecure, but also excited version of myself who four years ago dared to start this journey: "We made it".

Contents

List of Publications

The work introduced in this thesis is based on the following publications:

1. Carla Perez-Almendros, Luis Espinosa-Anke and Steven Schockaert. Don't Patronize Me! An annotated dataset with patronizing and condescending language towards vulnerable communities. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5891–5902, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.

2. Carla Perez-Almendros, Luis Espinosa-Anke and Steven Schockaert. Pretraining Language Models for identifying patronizing and condescending language: an analysis. LREC, Marseille, France, June 2022.

3. Carla Perez-Almendros, Luis Espinosa-Anke and Steven Schockaert. SemEval-2022 task 4: Patronizing and Condescending Language detection. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 298–307, Seattle, United States, July 2022. Association for Computational Linguistics.

4. Carla Perez-Almendros and Steven Schockaert. Identifying condescending language: A tale of two distinct phenomena? In Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 130–141, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics.

List of Figures

List of Tables

List of Acronyms

PCL Patronizing and Condescending Language

NLP Natural Language Processing

CDA Critical Discource Analysis

PoP Pornography of Poverty

ML Machine Learning

SoTA State of The Art

RNN Recurrent Neural Network

BoW Bag of Words

TF-IDF Term Frequency - Inverse Document Frequency

LM Language Model

SVM Support Vector Machine

BiLSTM Bidirectional Long-Short Memory

TL Transfer Learning

SA Sentiment Analysis

List of Abbreviations

COMMUNITIES

DIS Disabled

HOM Homeless

HOP Hopeless

IMM Immigrants

MIG Migrants

NEED In need

FAM Poor families

REF Refugees

VUL Vulnerable

WOM Women

CATEGORIES

UNB Unbalanced power relations

SHAL Shallow solution

PRE Presuppositions

AUTH Authority voice

MET Metaphors

COMP Compassion

MERR The poorer, the merrier

Chapter 1

Introduction

1.1 Background and Motivation

Nowadays, we live in a globalized world with increasing access to digital journalistic communication. In this context, how the media decides to cover news stories about vulnerable communities, might make an important social impact. For instance, the choice of a specific type of language when referring to an underrepresented group or a vulnerable situation, frames the story and positions both the author of the piece and the community they write about, by establishing (unbalanced) relations of power and privileges. Our research looks specifically at the use of Patronizing and Condescending Language (PCL) towards vulnerable communities, as a unique case of generally unintended harmful language widely used in the media. In the context of this thesis, PCL is defined as a type of discourse that shows a superior attitude towards others or depicts them in a compassionate way, raising a feeling of pity among the audience. [1]

[1] In our work we use the joined term *Patronizing and Condescending Language* because both *patronizing* and *condescending* are commonly used as synonyms in the literature, with works referring to one or the other indistinctly. Most dictionaries also treat them almost as interchangeable although some linguistic forums point to subtle differences in the definitions of both terms. The main difference lies in what actor takes the stage of the unbalanced relation, with *patronizing* usually referring more to the inferiority of the community or person object of the message, and *condescending* stating the superiority of its author.

Consider the following examples[2]:

> *"People don't understand the hurt, people don't understand the pain. I've read about women with their children sleeping in cars, sleeping in hotel rooms and it's criminal. If they're lucky and they come across COPE Galway and the ladies in Osterley, then there's hope."*

> *"December should serve as a time when we look with compassion at the fate of migrants, refugees and the internally displaced. It is especially a time when we must plan and increase resources for creative action."*

> *"Can't help if people want to flee a beggar country and take up citizenship of a good country so that their children become educated. If they live in hopeless for ever Sri Lanka they will end up as maids and servants in prosperous India and China."*

All three examples imply a position of superiority of the author regarding the person or community they are referring to, suggesting an imbalance in terms of power and privilege [52]. At the same time, the members of the vulnerable community are presented as victims, with the text explaining a painful, difficult or unfair situation, where the author and their audience are in a position to help or advise. This approach to vulnerability feeds what has been called *the Pornography of Poverty* [114], a communication style which explicitly depicts vulnerable situations in a compassionate tone to move its target audience to (charitable) action. The three examples express a dichotomy between a *victim* and a *saviour*, reinforcing the concept of the *(white) saviour syndrome* [9, 162], suffered by those in a more privileged position. Also from these examples, we can infer that the use of PCL is not always conscious and the intention of the author is often to help the person or group they refer to (e.g.

[2] These examples, as well as other examples that we will use throughout the thesis are taken from the DPM! dataset, a corpus of paragraphs extracted from news stories about vulnerable communities, that will be described in Chapter 3.

by raising awareness or funds, or moving the audience to action) [180, 105]. However, these superior attitudes [157] and the discourse of pity can routinize discrimination and make it less visible [115]. Thus, unfair treatment of vulnerable groups in mass media might lead to greater exclusion, inequalities and discrimination [116]. Moreover, due to its sometimes unconscious and generally well-intended nature, PCL is used by writers and accepted by their audiences with low defence, which further helps to enlarge distances, differences and inequalities between communities. This is why we believe that research towards the detection and understanding of PCL can help to build more responsible communication and, as a consequence, more responsible societies.

The discourse of condescension has been a topic of interest across a wide range of disciplines, including Linguistics, Politics, Journalism and Medicine [101, 59, 80, 28, 42, 118]. However, the rapid growth of digital communication calls for an automatic approach to PCL detection, such as the one Natural Language Processing (NLP) can provide. The study of unfair, ideological, offensive or misleading discourse has become an important and well-nourished topic of interest within the NLP research community. However, most works on this topic address messages with a flagrant and clear intention of harm or deception, such as hate speech [187, 188], offensive language [7], fake news [30] or propaganda [32, 33, 49]. More recently, researchers in NLP have also shown their interest in more subtle expressions of harmful language, such as condescension and other ways of expressing superior attitudes and power inequalities through language. For instance, in 2019 Wang and Potts [175] introduced the *Talk Down* dataset, which is focused on condescending language in social media. A year later and in the framework of this thesis, we released the *Don't Patronize Me!* (DPM) dataset [125], which is focused on the way in which vulnerable communities are described in news stories. Together with the dataset, we introduced the task of automatic detection of PCL towards vulnerable communities, which differs from previous attempts of addressing condescending language in two main aspects: 1) Our task focuses in indirect discourse, as it uses news stories covering vulnerable

communities and situations, which makes the condescension usually unconscious and with good intentions, and 2) we focus on the analysis of PCL towards underrepresented groups. Other recent works addressed some closely related aspects, such as how language conceals power relations [151], expresses authoritarian voices as empathy [195] or dehumanizes minorities [104].

However, PCL detection and categorization still present important challenges for the research community. Some of them are summarized as follows:

a. PCL tends to be a subtle and very subjective kind of language. The background of both the author and the audience of a message might influence in the assessment of a text as being condescending or not. Also, the good intentions behind a potentially condescending message might become an emotional barrier for some people to label it as condescending. These features of PCL would presumably make it harder to detect, both for humans and for NLP models.

b. PCL detection often seems to require commonsense knowledge and a deep understanding of human values and ethics [127], aspects where even the most advanced NLP models, such as Language Models, have shown their limitations [197, 85]. For instance, a person can use their knowledge of a specific sociopolitical situation or their understanding of morality to decide if a message contains PCL or not, whereas this would pose a significant challenge for NLP models.

c. PCL towards vulnerable communities is a complex phenomenon which encapsulates a broad spectrum of linguistic features, communicative purposes and ideological positions regarding potentially vulnerable communities. This suggests that there might not be a unique model to address all kinds of PCL and that, therefore, different approaches might be needed to detect and analyze different forms of PCL.

With these challenges in mind, we present our hypothesis and research questions in Section 1.2.

1.2 Hypothesis and Research Questions

Our research is based on a two-fold hypothesis:

1) Patronizing and Condescending Language can, to some extent, be automatically detected and categorized by Language Models, in spite of its subtle and subjective nature. However, its detection often requires commonsense reasoning, as well as world knowledge and an understanding of human values, which will pose a challenge for NLP models.

2) The analysis of PCL from an NLP perspective can help us to improve our understanding of the nature and features of PCL towards vulnerable communities.

In order to verify these hypotheses, the following research questions are addressed:

Research Question 1. How easy it is for human annotators to identify PCL towards vulnerable communities? Do human annotators agree in their assessments about the presence of this type of language?

Research Question 2. To what extent can Language Models identify and categorize PCL? Which NLP techniques are best suited to address this challenge?

Research Question 3. What does a model need to know to better identify PCL? To what extent would it need to understand human values?

Research Question 4. Can current State-of-The-Art NLP models effectively generalize to address the complexity of PCL?

1.3 Contributions

The primary contribution of this thesis is the introduction of a novel NLP task, namely the detection and categorization of Patronizing and Condescending Language towards vulnerable communities, as well as establishing the foundations for future work on this challenge with the introduction of quantitative experimentation and qualitative analysis. The specific contributions made through this research are as follows:

1. We created and released the Don't Patronize Me! dataset, which contains 14,299 paragraphs about 10 vulnerable communities extracted from media sources in 20 English-speaking countries or geographic areas. The dataset has been manually annotated to show 1) whether the paragraph contains PCL or not and 2) if it does, what category or categories of PCL are present in the text. We performed a qualitative analysis of the data and applied different NLP models for binary and multilabel classification in order to establish baselines for future work. The dataset is available under request for research purposes. This work was published at COLING 2020 [125].

2. We introduced a taxonomy of seven categories that can be used to express PCL towards vulnerable communities, based on previous research on the discourse of condescension made from other disciplines, such as Sociolinguistics, Cultural Studies, Journalism and Psychology. The taxonomy was included as part of the introduction of the Don't Patronize Me! dataset, published at COLING 2020 [125].

3. We organized a shared task on the Detection and Categorization of PCL towards vulnerable communities, which attracted the interest of 77 teams of NLP

researchers. From these participants, finally 38 teams submitted a paper summarizing their systems and contributions, which provided the community with novel approaches and insights for future work on this challenge. We proposed two subtasks: Subtask 1 presented a binary classification problem where participants' systems had to classify a paragraph as a positive or negative example of PCL; in Subtask 2, participants had to assign up to 7 labels for the categories of PCL that were present in a paragraph. This task was hosted at SemEval-2022 and a summary of the outcomes was published at NAACL 2022 [126].

4. We explored what kinds of pre-training could help a model identify and classify PCL. For doing this, we initially trained our models on a varied set of topics to later fine-tune them on PCL detection. We found that models' performance improves especially by transferring knowledge from data annotated with sentiment, harmful discourses such as hate speech or offensive language, or commonsense morality values. However, the limited gains acquired by this approach support the idea that PCL is a different and unique kind of language which requires different and specific approaches. This work was published at LREC 2022 [127].

5. We identified two main different phenomena in PCL, namely Linguistic and Thematic PCL. While Linguistic PCL relies on how a certain message is expressed, Thematic PCL is based on the message itself, what is being said. This latter type of PCL is harder to detect for NLP models if the training data does not contain the same condescending topics which can be encountered in the test data. We also showed that in the data available, some communities tend to receive condescension in a more linguistic way, while for others, PCL relies more on stereotypes or community-related themes. These findings were published at the Workshop on NLP for Positive Impact 2022 [124].

1.4 Thesis Structure

The remainder of this thesis is organized as follows:

- Chapter 2, **Background and Related Work**, provides an overview of the discourse of condescension and its potential impact among underrepresented groups. Furthermore, we describe current NLP techniques for text classification and review previous works on tackling PCL detection, as well as other related tasks.
- Chapter 3, **An Annotated Dataset with PCL towards Vulnerable Communities**, describes in detail the development of our dataset, including the curation, preprocessing and annotation of the data. We also introduce our taxonomy of PCL categories.
- Chapter 4, **Qualitative and Quantitative Analysis of the Data**, summarizes the findings of qualitative analysis on the Don't Patronize Me! dataset and presents baseline results for PCL classification and categorization using a set of popular NLP models.
- Chapter 5, **Pre-Training Language Models for Identifying PCL: An Analysis**, presents the different pre-training strategies we applied in order to transfer knowledge from related tasks to PCL detection. In this chapter, we collect the results of these experiments and perform a qualitative analysis to better understand the nature of PCL.
- Chapter 6, **Shared Task on Detecting and Categorizing PCL**, analyses the contributions of some of the best-performing teams which participated in our shared task on PCL detection, and compares the results with new baseline results on the same data setting. We analyse the models' performance and the findings and insights by the participating teams and draw conclusions for future work.

- Chapter 7, **Identifying Condescending Language: A Tale of Two Distinct Phenomena?**, describes further experimentation and analysis on the data based on the findings explained in Chapter 6, which led to the identification of two different phenomena in PCL, namely Linguistic PCL and Thematic PCL.
- Chapter 8, **Conclusions and Future Work**, concludes this thesis by reviewing how the results and findings reached during our work address the hypothesis and research questions initially stated. We reflect on how this work is a valuable knowledge contribution to the research community and for society in general and suggest roads for future work in this area.

1.5 Summary

In this chapter, we have introduced the main topic of this thesis, namely the automatic detection of Patronizing and Condescending Language towards vulnerable communities. We have also justified the motivation behind our research topic and briefly introduced some related works that others have conducted before us. Moreover, we have discussed the challenges that still remain, our hypothesis and research questions, as well as the main contributions derived from our research. The next chapter will provide an extensive discussion of the related work in order to put our research in context, before deepening into the technical contributions. We will furthermore describe the NLP techniques that we have relied on in our work.

Chapter 2

Background and Related Work

2.1 Introduction

This thesis is focused on the automatic analysis of a specific type of biased discourse, namely Patronizing and Condescending Language towards vulnerable communities. The discourse of condescension, as well as the relationship between power and language, has been widely studied from different disciplines, such as Politics, Sociolinguistics and Media Studies [101, 59, 81, 28]. Many researchers in NLP have also focused on studying biased language. However, traditionally, most studies have focused on flagrant phenomena, such as hate speech or offensive language [187, 188, 7], but have neglected discourses which would undermine others in a more subtle way, such as PCL. Nevertheless, a number of recent works have shown an interest in the analysis of PCL [175, 125], and related discourses [151, 104]. In this chapter, we put this thesis in context, reviewing the existing literature on the research areas and approaches related to our work. Specifically, in Section 2.2 we explore works from other disciplines to learn more about Patronizing and Condescending Language, as well as other kinds of discourses with similar effects on society. In Section 2.3 we review the main NLP approaches to text classification used in our experiments. Last, Section 2.4 offers an overview of previous works on the automatic detection of different types of human bias in text. Here, we also review the latest approaches to detect the expression of social inequalities in

text, such as the use of condescending language.

2.2 Patronizing and Condescending Language

2.2.1 Brief History of Condescension

The discourse of condescension has been widely studied throughout history, attracting the interest of scholars from very different fields, such as Language Studies [101], Sociolinguistics [59], Politics [80] or Medicine [92].

Language has been used since the first civilisations of humanity to establish (unbalanced) power relations and to maintain the status quo. The discourse of condescension arises as an inherent part of that relation between language and power, as it implies the participation of (at least) two actors in the communication process, one in a more privileged situation (source) than the other (either the receiver or the object of the message).

Being condescending always connotes a difference in the status, class or situation of the actors of the communication process. However, the meaning and social use of condescension have changed over history. In his work on The Failure of Condescension [155], Daniel Siegel analyzes the evolution of the social acceptance of a condescending discourse through its use in literature. He claims that until the 18th Century being condescending towards someone was received with gratitude and admiration, as it was a concession for those in higher positions to care for those inferior to them. However, in the early decades of the Victorian era (1837-1901) and coinciding with some social revolutions (e.g. the voting reform or the idea that vulnerable people were not *victims* and that any person was able to help themselves [159]), condescension begins to have negative connotations. As Siegel explains it, to condescend someone was not seen anymore as a renounce to oneself, but as a way of *showing* that renunciation; it was not seen as a way to help others, but

to demean them and their capabilities for one's own gain. This description of the concept of condescension is still valid today. However, besides its negative connotations, Patronizing and Condescending Language is used every day in a variety of contexts, such as the health domain [143, 48], in inter-generational communication [58, 83, 60, 13], in political discourse [82], in social media [175] or in news articles [125], to name just a few.

2.2.2 PCL towards vulnerable communities

PCL towards vulnerable communities, as we study it in this thesis, is a subtle and subjective kind of language where the author of a message expresses a superior social, moral or intellectual attitude or position towards a third person or community, often presenting them as victims or as unable to overcome their own situation. However, the use of this language, especially when targeted to vulnerable communities is often unconscious and well-intended [180, 105]. An author might use PCL while trying to help a community or individual, raise their voice for them or move their audience to (charitable) action. Nevertheless, the potential impacts of PCL can be very harmful. Based on research on Sociolinguistics and cultural studies, we collect here some of the potential harms of PCL:

- it fuels discriminatory behaviour by relying on subtle language [104];
- it creates and feeds stereotypes [51], which drive to greater exclusion, discrimination, rumour spreading and misinformation [116];
- it establishes, consolidates and implements power-knowledge relationships [52] by positioning one community as superior to others;
- it usually calls for charitable action instead of cooperation, so communities in need are presented as passive receivers of help, unable to solve their own problems and in need of external help to solve them;

- it reinforces the dichotomy of a *saviour* and a *helpless* victim [9, 162];
- it tends to avoid stating the reasons for very deep-rooted societal problems, by concealing those responsible or even, in some cases, by apportioning blame to the underprivileged communities or individuals themselves [168];
- it proposes ephemeral and simple solutions [29], which oversimplify the wicked problems [72] vulnerable communities face;
- it contributes to the "distorted and stereotyped representation" [24] that vulnerable communities and underrepresented groups frequently receive in the media.

In summary, PCL routinizes discrimination [115] and makes it less visible, making it more difficult for vulnerable communities to overcome difficulties and reach total inclusion [116], especially when widespread by the media. Therefore, the news discourse becomes a subtle but consistent player in the maintenance and legitimization of social inequalities [168].

In order to understand better the concept of PCL towards vulnerable communities, we should not disassociate it from other closely related concepts that construct, together, a system of inequalities supported by language. We review some of these concepts in the next subsections, namely Discourse and Power, The Practice of *Othering*, The Pornography of Poverty and The Coverage of Minorities in the Media.

2.2.3 Discourse and Power

The relation between discourse, power and social inequalities has been widely studied in Social Sciences, with scholars such as Pierre Bourdieu [16], Michel Foucault [53][1] and Jürgen Habermas [67] as main representatives of this research area. Their

[1] Originally published in 1972

works motivated the appearance of the *Critical Discourse Analysis (CDA)*, an interdisciplinary approach that emerged from Critical Linguistics in the 1970s that looks, among others, at the relation between power and language, and analyzes how discourse expresses social hierarchies and inequalities [46, 45, 181, 182, 28, 80, 54].

Previous works on CDA highlight how language can reinforce inequalities and exclusion and perpetuate unbalanced power relations and privileges, which are distinctive features of PCL. For instance, in his work *Critical discourse analysis and the discourse of condescension*, Huckin [80] states some of the linguistic techniques used by authors of written text to express a condescending treatment. Next, we present some of the most relevant for our work on PCL, together with examples extracted from the *Don't Patronize Me!* dataset, which will be presented in Chapter 3:

- **Classification** refers to the selection of a specific word, name or label to describe a situation or a community. Consider, for instance, a news article's headline calling the Rohingya migrants "The Unwanted". The selection of words here is already showing certain consideration towards the community.
- **Connotation** is used when the selection of words made by the author of a condescending message carries more meaning than the one stated in a dictionary. For instance, the reference to *Western values* is inherently opposing these to other sets of values, which in some contexts might imply an idea of superiority or condescension.
- **Metaphors** are also used to express something in a different way, and are helpful as a means for euphemisms or to give a poetic touch to vulnerability. For instance, expressions like *restore the dreams of someone* use metaphors to express condescension.
- **Presuppositions** are very common in PCL and are used when the author states as a fact something that is not known. For example, an author stating "they will end up as maids and servants" is making a presupposition.

- **Modality** refers to the use of modal verbs, which are widely used in condescending language and project authority voices and attitudes. For example, the sentences "we *must* help them" or "they *should* know better" reflect a condescending treatment.

- **Transitivity** refers to who takes the roles of agent and patient in a sentence and might be one of the most distinctive linguistic features of PCL. Sentences like "The photo of a Hyderabad traffic policeman feeding an elderly homeless woman has gone viral" or "the ladies had the chance to share their compassion for those in need" show a clear difference between the agent of the action, who is presented as a *saviour* of those in need, who are *victims* and passive receivers of their help.

- **Deletion** or the deliberate omission of relevant information can be also used to express power relations. For example, the *agentless passive*, or the construction where the author omits the agent of the action, can be found in the following sentence: "Lima is the fifth most dangerous city for women in the world behind Cairo, Karachi, Kinshasa and Delhi. There are 30 attacks on women a month and 10 are killed".

However, it is important to highlight that, as PCL, the expression of power inequalities through discourse is not necessarily intended or even conscious. It emerges just as an expression of the *status quo* that the powerful, who traditionally enjoy the privilege of public discourse, perpetuate through common-sense assumptions implicit in any linguistic interaction. As Fairclough [45] states, these assumptions constitute the ideology of the author and thus, unbalanced power relations are reinforced simply by reproducing the ordinary ways of behaviour towards others less privileged.

2.2.4 The Practice of Othering

Another linguistic feature that helps us identify PCL towards vulnerable communities is the distance that an author establishes between themselves and the community they are referring to. With this treatment, the unprivileged group and their circumstances are referred to as something alien to our experience. They become *the others*, different from *us*. According to Brons [17], the notion of *othering* emerges from Hegel's *Master-Slave dialectic*[2] and grows in Post-colonial Studies and Feminist Theory before spreading to other research areas, such as Psychology [93], Healthcare [23] or Cultural Studies [37].

The practice of othering has served to distance communities and to create and reinforce identities, both self and aliens'. It constitutes a pillar for racism, discrimination, exploitation, wars or genocides [17], as the moment we do not identify a group of people with *us*, we allow ourselves to treat them and consider them differently [23]. For instance, the othering is behind the creation of the concept of *Orient* as the antithesis of *Occident* in a context of European or *West* Imperialism [148].

PCL also relies on othering vulnerable communities, in this case not for justifying an offensive treatment, but for reinforcing the difference, the inequalities and the power structures that put the author of a message in a position to condescend others. Through this practice, the roles between both sides in a communication process are assigned, with one being in a more powerful position than the other. In addition, and according to Thomas-Olalde and Velho [166], "the formula of constructing a positive self-image via the construction of a negative (or vulnerable) image of the other" lies behind the creation of stereotypes, another distinctive feature of PCL.

[2] Published in his book *Phenomenology of Spirit* in 1807

2.2.5 The Pornography of Poverty

The Pornography of Poverty (PoP) is a term used by development practitioners to refer to a communication strategy that exploits either images or descriptions of poverty and other tough situations to reach an often privileged audience and move it to charitable action (e.g. make donations) [114, 131]. Such strategy has been proven to work well for NGOs and news media, which have traditionally used it to achieve their objectives, such as raising funds or engaging readers. However, the harmful effects of this practice have also been stated by many [119, 131, 114]. On the one hand, the continuous exposure to dramatic images and stories seems to make audiences get used to them, relaxing the feeling of pity and guilt and, as a consequence, the movement to action. On the other hand, there is, at least, an equally harmful consequence of the Pornography of Poverty and this is the effect that it has on the object of such messages. Through the PoP, vulnerable communities are presented as helpless, passive objects of others' actions, victims awaiting a saviour [131]. According to an article by Anne Buchanan in a publication of the Canadian Council for International Co-operation issued in 2002[3], with the PoP NGOs (and media sources) practice a "yellow development", where the sensationalism of the story is more important than the ultimate goal of equality and social justice. The use of the PoP, therefore, poses many ethical issues. In the first place and according to Oliver [119], the objects of such messages are hardly ever consulted about the portrayal the media offers of them. Second, we should ask ourselves (as the privileged side of this story) if that kind of communication works because it appeals to our solidarity or because it reinforces our feeling of superiority. Also, as PCL, the PoP ignores the responsibility of more powerful communities over inequalities, portrays a narrowly framed picture of a bigger and extremely complex reality and creates and reinforces stereotypes of need and incapacity.

[3] Anne Buchanan, "Beyond Stereotypes: Seeking New Images," Au Courant 11, no. 1 (spring 2001), 4–6. https://ceim.uqam.ca/db/IMG/pdf/004_au_courant_spring_2002_en.pdf (Last access May 2023).

2.2.6 The Coverage of Minorities in the Media

Previous works on the relation of power and discourse also draw our attention to the influence (voluntary or not) that public discourse has on society and how it helps to construct a specific and inherently biased image of a certain situation or community in the mind of the audience. The media, as a speaker of public discourse, has a great responsibility in the construction of these mindsets, for instance by using recurrent themes and stereotypes in the coverage of minorities [80]. According to Van Dijk [168], the discourse of media contributes to the construction of ethnic identities in society, which contributes to the maintenance of inequalities and unfair treatment towards vulnerable communities. Along this direction, Huckin [81] studied the treatment of homelessness in the US in 1999. He collected a corpus of 163 newspaper articles and editorials which mentioned the keyword *homeless* and analyzed, among others, the more recurrent themes and stereotypes related to this community. Although the objective of his analysis was to detect a framed picture of the homelessness reality, we also identify a condescending one. For instance, he shows that the analyzed data includes "desire of independence" or "lack of life skills" as common themes when referring to causes of homelessness. Moreover, the theme "bad grooming" is highlighted as one effect of homelessness, and "religious support", "food donation" and "donated clothes" are common themes in the discussion of public responses, which represent shallow and ephemeral solutions for a structural, deep-rooted problem, and thus again reinforce the charitable, *saviour-victim* treatment of a community. Using a similar approach, Díaz-Rico [39] analyzed 93 articles about Mexican immigrants from the Los Angeles Times, published in 2010. She claims that the selection of topics and themes is the most important aspect of Journalism and that newspapers use the drama of a story to gain attention from their audience. Although the language and topics she analyzes in this work are often openly discriminatory and offensive, she also finds expressions that, through rhetorical figures, connotation and semantic selection, reinforce power relations and

inequalities (e.g.“help new arrivals get on their feet”, or "ballot crusade").

2.3 NLP for text classification

2.3.1 Introduction

Natural Language Processing is the branch of Artificial Intelligence that focuses on the automatic processing, analysis, generation and understanding of human language. Some of the most popular NLP tasks include Text Classification, Machine Translation, Question Answering or Summarization, to name just a few. In this thesis, we focus on text classification, a downstream task in NLP that tries to build models which are able to automatically classify a given textual input into a set of different classes. Specifically, we work in a supervised learning setting, where the training data provided to the model contains labels which assign one or several classes to each input sequence. By training on that labelled data, the expected outcome is that the model learns to associate specific features with each one of the classes, being able afterwards to classify previously unseen and unlabeled textual inputs.

2.3.2 Brief History of NLP

Since the 1950s, different approaches have helped to improve the performance of models on the task of automatic text classification. Next, we briefly introduce the three main stages in the history of NLP.

Symbolic NLP (1950s-1990s). The earliest approaches to NLP consisted of rule-based programs. Through a more or less complex system of rules, these programs were able to simulate natural language understanding, or to interact to a certain level of satisfaction with humans. Rule-based programs are still used nowadays, in

spite of presenting certain limitations, such as the lack of sustainability or scalability, as the models would need new rules to address new challenges. However, some valuable resources developed during this stage of symbolic NLP, such as ontologies, are still widely used in the field.

Statistical NLP (1990s-2010s). The arrival of the *statistical revolution* [86, 144] in the late 1980s and early 1990s, together with the rapid growth of computational power, gave way to a new paradigm, i.e., Statistical NLP. The new approach developed Machine Learning (ML) with statistical linear models, such as SVMs [31] or Logistic Regression, which would *learn* from large corpus collections. After the birth of the Web 2.0[4], these large amounts of data were more easily available due to the exponential growth of digital communication and the democratization of the Internet.

Neural NLP (2010s-today). During the decade of the 2010s, neural networks introduced new State of The Art (SoTA) results in ML in general and NLP in particular [61, 63]. Neural networks process sequential data through a certain number of connected layers which act as non-linear classifiers. Since the seminal work by Vaswani et al. [171] with their paper *Attention is all you need*, the *transformers* architecture took the stage in NLP. This novel approach has achieved SoTA results in most NLP tasks, surpassing Recurrent Neural Networks (RNNs), such as LSTMs, which were the most preferred models until the moment. Unlike RNNs which would process data sequentially, transformers are able to process the entire input at the same time, thus considering all the context for each word in a given sequence. This allows parallelization, which leads to faster training processes. In turn, this facilitated the apparition of large pre-trained language models.

[4]The term was first introduced by Darcy DiNucci in 1999 in his article Fragmented Future: https://www.webdesignmuseum.org/web-design-history/web-2-0-1999 (last access May 2023), and later popularized by O'Reilly in 2004[120]

2.3.3 Data representation

In most NLP approaches, textual data needs to be presented as a vector or a numerical representation of the words contained in a textual input, so that a computational model can process it. In this thesis, we use three different types of representations in our experiments: Bag of Words vectors, word embeddings and contextual embeddings.

Bag-of-Words (BoW). The Bag-of-Words model is one of the simplest ways of representing a textual input. It encodes how many times each word from a given vocabulary is present in a given text. The vocabulary can be given to the model as an external set of words (for instance, a list of adjectives) or can be extracted from the analyzed corpus using different techniques. One of the most popular approaches is *Term Frequency- Inverse Document Frequency* or *TF-IDF*, which weights the importance of each term by comparing the frequency of that term in a given document with its frequency in the whole corpus. In this way, a *k* most-important-words vocabulary can be extracted from the corpus with those words which are more relevant for a given classification task. BoW representations do not consider the order of the words or their relations with other terms, just whether they are present in the input text, and usually generate very sparse vectors.

Word embeddings. A word embedding is a way of representing the meaning of a word as a numerical vector in a vector space. Word embeddings emerge from distributional semantics and the intuition that words with similar meaning should be close to each other in a vector space, or as the popular cite by Firth says: "You shall know a word by the company it keeps" [50]. Thus, these representations use neural networks trained on large collections of data to assign each word a numerical representation or vector that reflects its meaning. Word embeddings are known to represent very well semantic and syntactic properties of language and have been

widely used in many downstream NLP tasks, especially during the 2010s. In this thesis, we experiment with word embeddings for our PCL classification task. Specifically, we apply one of the most popular models of word embeddings, namely Word2Vec [107], which will be explained in more detail in Chapter 4.

Contextual embeddings One of the main limitations of word embeddings is that they represent each word with a unique vector, without considering its context. Thus, some semantic properties, such as polysemy, can not be captured by these models [97]. This is solved with the apparition of contextual embeddings, which will look at the entire input at once, therefore considering the context for each one of the generated embeddings. These contextual embeddings are trained mainly with Language Models (LM), which use the transformer architecture to learn word representations. These representations are nowadays used on most downstream NLP tasks, including text classification. Their associated LMs, such as BERT, constitute the most popular approach in the majority of NLP challenges, achieving SoTA results in most of them. We will talk more about LMs in the next section.

2.3.4 Popular Approaches to Text Classification

In this subsection, we review some of the most popular approaches for text classification. We only consider the models which we apply to our experiments, namely Support Vector Machine (SVM); Recurrent Neural Networks (RNN), specifically Bidirectional Long-Short Term Memory neural networks (BiLSTM), and Language Models, built over a transformers architecture.

Support Vector Machine. SVM is a supervised Machine Learning technique that tries to separate instances from different classes by maximizing the margin between them. A SVM model projects the input instances in a vector space and *draws* a

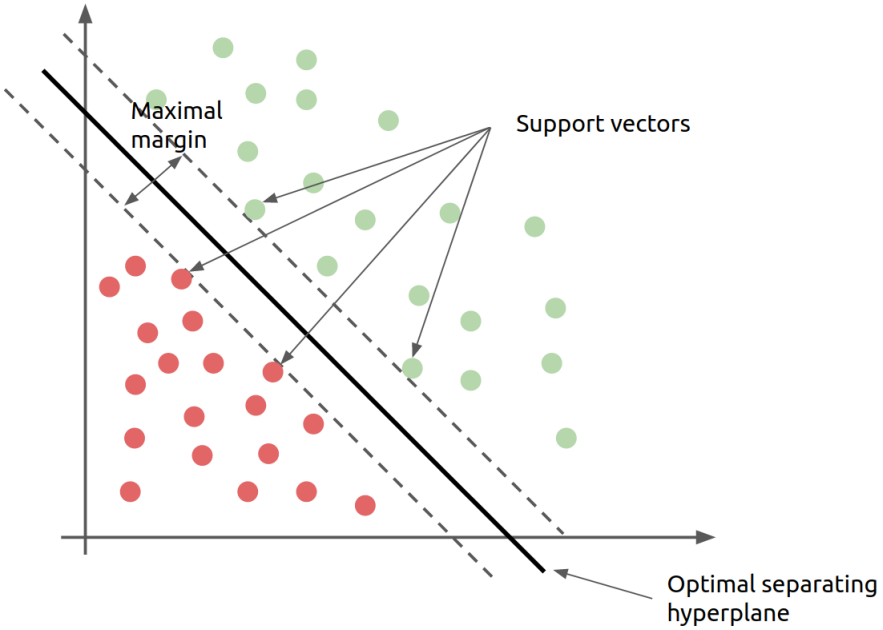

Figure 2.1: Representation of a linear SVM

hyperplane able to divide the instances of the different classes. The test instances are then projected into the same space and the model will assign them a class depending on where in the map they fall. SVMs can also handle non-linear problems, by projecting the input instances to a higher dimensional space and modifying the kernel configuration. Figure 2.1 shows a representation of a linear SVM.

Recurrent Neural Networks. Neural networks are another Machine Learning technique widely used in supervised learning. They take inspiration from how the human brain works, trying to replicate the synapses in our brain with the connection between artificial neurons. A neural network takes a sequence as input data, which passes through a set of hidden layers, each one with a specific number of connected neurons or nodes, until reaching a last output (or classification) layer, which will assign a class label to the input data. Neural networks are usually initialized randomly, by assigning random weights to the connections between nodes. These weights are readjusted during the training process, by comparing the predicted label with the

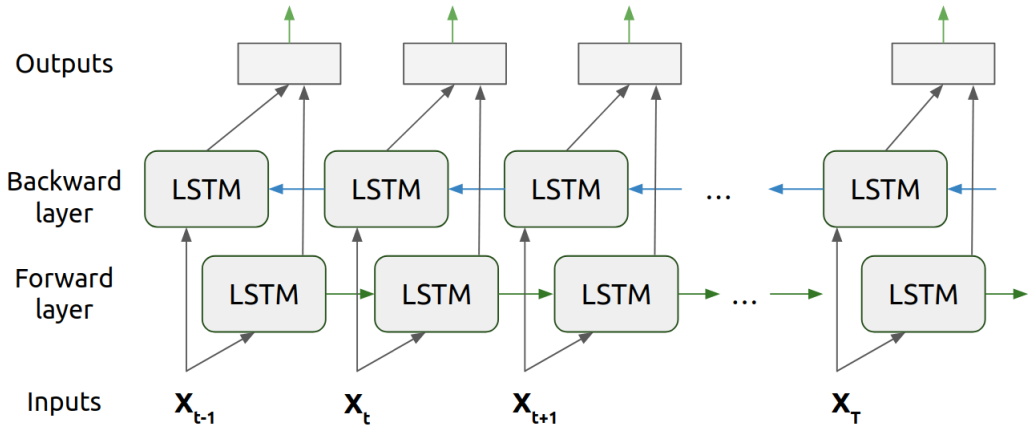

Figure 2.2: BiLSTM architecture

correct label [61]. Recurrent Neural Networks [44] are good at capturing not only syntactic and semantic properties, but also structural features of the input data, as they consider the context of each word. This can be done in a feed-forward way or, as is the case of BiLSTM, in a bidirectional way [74, 153]. However, neural networks only look at each word's context by considering the previous and the next word in the input sequence, which might still miss relevant contextual features of the data. Figure 2.2 shows the architecture behind a BiLSTM neural network.

Pre-trained Language Models. Pre-trained Language Models constitute the most popular approach in NLP and are achieving SoTA results in most NLP tasks, including automatic text classification. LMs are based on the transformers architecture [171], which, for the first time, allowed automatic models to process all the words of an input sequence at the same time, using a novel attention mechanism to provide context for every position of the sequence. On the one hand, this leads to the creation of contextual embeddings, which are able to represent the meaning of each word in context and which are inherently linked to LMs. On the other hand, the capacity of processing the whole sequence at once reduced training times, allowing these models to be pre-trained on large corpora, which in turn allows them to cap-

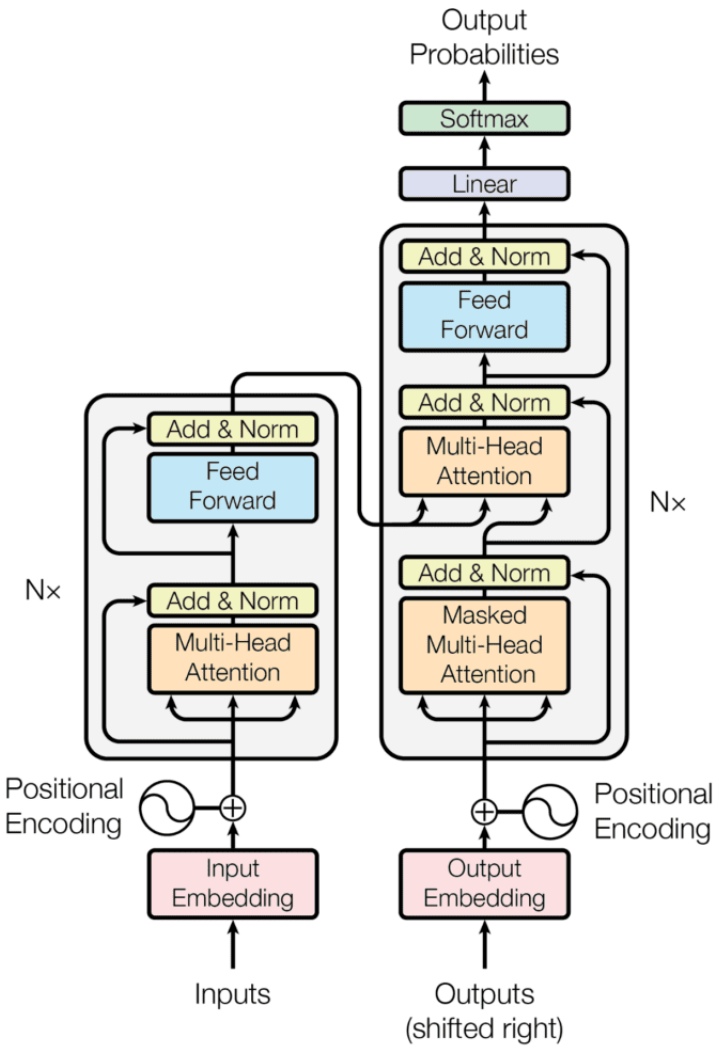

Figure 2.3: Transformers architecture, extracted from Vaswani et al. [171]

ture more subtle or implicit features of language. BERT(Bidirectional Encoder Representations from Transformers) [38], is one of the most popular LMs, from which many others have been developed afterwards. Figure 2.3 shows the transformers' architecture.

One of the most popular approaches using LMs consists of initializing the model from the pre-trained version, e.g., pre-trained BERT embeddings, and fine-tuning those initial representations in the target task intended to be solved. By fine-tuning, the weights for each word are updated with new specific knowledge about the task (e.g.,

linguistic features). If the training data is labelled, we can also train a classification layer which will allow the model to classify new data. For instance, if we want to fine-tune a BERT model on PCL classification, we will fine-tune both the BERT model and the classification layer at the same time.

2.3.5 Transfer Learning

Transfer learning is a ML technique that tries to improve the performance of automatic models by transferring the knowledge from a similar task to the target task, with the expectation that the transferred knowledge will help to solve the final challenge [196]. These related tasks are also called auxiliary or source tasks. For instance, a model aimed to perform well in classifying PCL might need to learn before about stereotypes or moral values. This knowledge, presumably, will help the model to identify whether a specific text is condescending. Approaches using transfer learning would first train a model in one or several auxiliary tasks. The resulting model would then be trained and evaluated in the target task. Transfer learning can be applied either to provide the model with related knowledge that might help solve the task, or to provide extra information if there is scarce availability of training data on the desired domain [179].

However, the knowledge transferred from other tasks does not always help to improve the performance on the target task. The ***negative transfer*** [66], or a loss in the performance of a transfer learning model, might occur due to several reasons, such as the limited relevance of the source task for the target domain, or the limited capacity of the model for transferring the appropriate knowledge across domains [196].

The ***catastrophic forgetting*** [102, 142] is another undesired consequence of some approaches of transfer learning, such as fine-tuning LMs. In this process and as we already saw before, the weights of the LM change to adapt to the new task,

which might make them *forget* some of its previous knowledge. Considering that these models have been pre-trained on vast amounts of data, missing part of that knowledge might lead to a poorer performance on the target task, even when the knowledge transferred by the auxiliary task is appropriate.

According to Zhuang et al. [196] transfer learning (TL) approaches can be categorized from different aspects, such as the problem we are trying to solve or the solution we are seeking.

On the one hand, considering the problem categorization, we can classify TL as follows:

Depending on what labels are available:

- Transductive, if we only have the labels for the auxiliary or source task.
- Inductive, when the labels for the target task are available.
- Unsupervised, if no labels are available for any of the two tasks.

Depending on the consistency between auxiliary and target data domains:

- Homogeneous, when both tasks share the same domain or the differences between their feature spaces are just marginal.
- Heterogeneous, when the domains of the source and target tasks are different and the representation of a word in one task does not necessarily correspond to the same word's representation on the second task.

On the other hand, looking at the solution categorization, TL can be:

- Instance-based, which consists of assigning different weights to the instances in the source domain in order to adapt them to the target task.

- Feature-based, which transforms the representation of the original features into a new representation, more suitable for the target domain.
- Parameter-based, when the knowledge is transferred at model level.
- Relational-based, which transfers the logical relations learned from the auxiliary task to the target task.

In this thesis, we experiment with inductive, heterogeneous, parameter-based TL. Specifically, we use LMs to extract knowledge from several related tasks which can be later applied to PCL detection.

We use two different approaches, namely full fine-tuning and the use of adapters [75], which are introduced as a means to avoid catastrophic forgetting. We will talk about these approaches in Chapter 5.

2.4 Bias in NLP

2.4.1 Sources of bias in NLP

Bias in NLP can be defined as the (lack of) fairness of NLP models regarding their performance on different types of data or populations. Bias might be introduced during the process of selecting and processing data, designing the research process or building the models.

Hovy and Prabhumoye [76] outline five different sources of bias in the NLP process, which we briefly explain as follows:

- **Bias from data selection.** The moment a researcher selects some data to develop their experiments, they are introducing bias. As an example, consider the source of the data. Choosing data from media sources will present the models

with a language that differs from that used on Tik-Tok, a social media platform mainly used by teenagers. Also, the demographics of the authors will be different, as it will be the purpose of the message. A model trained on the first type of data, therefore, will probably perform poorly on the second type of data. Even in a model meant to work only with data from a specific source, such as media, the data selection will not represent all the data available, therefore the selection will again introduce bias such as language or political stance.

- **Bias from data annotation.** The annotation process will also introduce bias in the process. The demographics, background or values of the annotators will inherently bias their annotation [152], especially if the labels considered on the task are subjective or rely on previous knowledge. For instance, if the task implies assessing offensive language towards women, a female annotator and a male annotator might introduce different types of bias. Furthermore, a mismatch between the annotator's and the author's social or linguistic rules might lead an annotator to interpret a message differently from the way in which it was intended [150].

  For our dataset, we will address these two kinds of data-related biases in Chapter 4, following Bender and Friedman [10]'s work on mitigating bias in NLP systems.

  The remaining three sources of bias following Hovy and Prabhumoye [76]'s work are not the focus of this thesis, but we acknowledge their influence on the performance of NLP models. These are as follows:

- **Bias from input representations.** Embeddings carry bias and stereotypes from the data they are pre-trained on [22, 19, 163]. Detecting and mitigating these biases is an important area of interest for many researchers, both for word embeddings [14, 99, 62] and contextual representations [192, 8]. These biases are sometimes amplified in cross-lingual settings [193].

- **Bias from models.** In addition to the biases contained in the data and car-

ried by word representations, Machine Learning models introduce a new issue called bias overamplification [68, 173]. Overamplification happens when machine learning models not only reproduce, but amplify the biases encountered in data, making certain predictions at a higher rate than expected based on the data statistics. For instance, according to Zhao et al. [191], given a dataset where the activity of cooking was 33% more likely to be associated with women, a trained model would amplify the probability of predicting the class *women* to 68% on the test set. Bias amplification by ML models might lead to perpetuating undesired stereotypes [190, 40, 191] or to achieving unequal performance depending on the analyzed population [20].

- **Bias from research design.** The way in which a researcher designs their work will also introduce biases in the process. For instance, the language of the analyzed data, the geographic origin and demographics of their authors, the topics we choose to do our research on or even the composition of research groups will bias the model and its results [76].

We will cover the biases introduced in our research in the following chapters, especially in Chapter 3, where we introduce our dataset.

2.4.2 NLP to detect human bias

We consider that being aware of and transparent with the biases we introduce in our research is a social responsibility for researchers. In addition, we think bias has a different dimension in NLP, which is where our work fits in. NLP can help to detect and mitigate human bias in order to help us achieve more responsible communication.

NLP has been targeting biased language for a long time, and tasks such as sentiment analysis [133, 134, 146, 122, 6], hate speech [7] and offensive language detection [187, 188] have become well-established challenges in the community. PCL

is an equally harmful but subtler kind of language which has not yet received as much attention as other more aggressive, flagrant or explicit phenomena. However, we consider that the advances in the detection of other human biases will also help to detect condescension. Next, we briefly introduce some of the main NLP topics that might also help with our target task:

Abusive language. Online abusive language has been a central topic of NLP research for a long time, whether it is generalized or directed towards a specific person, explicit or implicit [177]. One of the most studied types of abusive language is hate speech, which attacks a group or an individual based on perceived aspects of identity (e.g. sexual orientation, ethnicity or ideology) [176] and which according to Yin and Zubiaga [185] can be particularly harmful for marginalized communities. Although hate speech and offensive language have been used as synonyms in the past [176, 64], there are important differences between them. For instance, hate speech, as a type of abusive language, has a clear intention of damaging others, whereas the offensiveness of a message might be more subtle and depend more on the receiver's response [185]. Therefore, hate speech should not be mistaken with PCL, as the intention of the message is completely different. However, a condescending message will be necessarily offensive if the condescension is perceived. In spite of the aforementioned differences, the target of any harmful message is usually the same, those more unprivileged or underrepresented communities. For this reason, we consider that learning about hate speech and offensive language might contribute towards a better (automatic) understanding of PCL.

Stereotypes and Rumour Propagation. Messages about vulnerable communities or underrepresented groups often feed stereotypes and assumptions which contribute to discriminatory behaviour and the creation and propagation of rumours. This creates a potentially harmful collective mindset regarding underrepresented communities, feeding misunderstanding, hate and exclusion [15]. In the last years,

researchers in NLP have tried to detect the presence and propagation of these phenomena [112, 198] or the veracity of them [36, 65].

Stereotypes, as well as rumours, are not necessarily ill-intended. In fact, rumours could be the expression of wishful thinking, although they can also feed fear or hate [90]. Stereotypes might also try to highlight positive characteristics. For instance, stereotypes such as *resilient*, *strong* or *brave* might sound like familiar attributes assigned to vulnerable groups in the media. However, the spreading of a biased assumption or unverified information about a community does not help mutual understanding and inclusion. A biased consideration of *others* might lead to discriminatory behaviour (or language) towards them, either more flagrant and aggressive, such as hate speech, or more subtle and well intended, such as condescension.

Human values, sentiment and emotion detection. Researchers in NLP have also tried to incorporate social values and emotions into their models in order to better reproduce human behaviour. For instance, sentiment analysis (SA), which pursues the automatic classification of an opinion's polarity, has become one of the most popular NLP tasks both in research and in industry [133, 134, 146, 122, 6]. Over the last years, the task of SA has evolved from a simple, binary classification task to more sophisticated and fine-grained analysis [95], which has allowed the NLP community to face different and more complex challenges.

The task of emotion detection [110, 25, 154] constitutes a different scenario from SA, as the focus of the analysis falls on the subject, instead of the object of the opinion [113]. The classification task becomes more complex, as it is generally a multiclass problem [132, 55] and the emotions might be expressed in a more subtle way [103]. As a third step in complexity and abstraction, we might find the task of human values detection. Researchers such as Rezapour et al. [145] or Hendrycks et al. [73] have provided the community with resources such as lexicons and annotated datasets which are aimed at training NLP models to understand human values.

Although these approaches try to make NLP more human, they do not come without ethical concerns, as introducing social values will inherently introduce social biases. Moreover, most models try to focus on universal principles, however, every situation might involve specific values [165], which makes it more difficult for models to generalize [96].

PCL detection. The types of discourses previously discussed in this section are an example of the traditional focus of the NLP community when studying harmful language. As Banko et al. [4] present in their taxonomy of harmful content online, harmful language has been traditionally seen as a more or less explicit, aggressive and flagrant phenomena. However, most of these works only refer to speech that purposely intends to harm others, leaving out other kinds of unintended or at least more subtle but equally harmful ways of undermining language, as is the case of PCL.

Recently, however, some work on condescending language has started to appear. For instance, Wang and Potts [175] introduced the task of modelling condescension in direct communication from an NLP perspective, and developed a dataset with annotated social media messages. They reckon the difficulty of detecting condescension both for automatic models and humans, who are sometimes unaware of being the target of condescending treatment. Moreover, one of their conclusions reinforces the idea that context is often needed to classify a message as containing condescending language. Although their work focuses on texts which are intently condescending, they acknowledge the existence of messages that can condescend in an unconscious way. However, the damage these unconscious messages can cause is hardly mitigated by the lack of intent.

Although the focus on condescension is a recent phenomenon in NLP, other works have focused on closely related topics before. For instance, Sap et al. [151] discuss the social and power implications behind certain uses of language, an important

concept in the unbalanced power relations that are often present in condescending treatment. In their work, they highlight that the power implications of certain discourses are reflected in the implicit meaning of what is said, and because of this, NLP models often find difficulties to detect such language. Moreover, they release the Social Bias Inference corpus, with annotations of social implications towards specific groups, which is aimed at training models on identifying social biases. However, they recognize the need of injecting commonsense reasoning into NLP models in order to help them understand these bias frames.

Also related to unfair treatment of underprivileged groups, Mendelsohn et al. [104] analyzed, from a computational linguistics point of view, how language has dehumanized minorities in news media over time. They point out the power that media has in creating a specific mindset towards underrepresented groups and how implicit dehumanization of some communities might lead to discrimination and hate. The concept of *othering*, as we saw before, is also related to the dehumanization of vulnerable communities and to condescension. Authors like Burnap and Williams [21] and Alorainy et al. [3] have addressed the detection of this harmful practice.

Other authors such as Ortiz [121], Price et al. [136], Gilda et al. [57], Han and Tsvetkov [69] have also considered condescension as a type of subtle but harmful language present in *unhealthy* conversations.

2.5 Summary

In this chapter, we have reviewed the existent literature on condescension and other closely related topics that reflect the relation between language and power and which are relevant for understanding PCL. We have also introduced the field of NLP for text classification, as well as some of the most popular techniques which have been used in our experiments. Last, we offered an overview of those NLP tasks and works which have focused their attention on the detection of human biases, with a last

section dedicated to efforts on detecting more subtle but equally harmful types of discourse, such as condescension. With this overview of the related work, we have put our thesis in context and have justified the need for our research. In the next chapters, we will introduce the motivation that has moved our work, the experiments conducted and the results obtained in the framework of this thesis.

Chapter 3

An Annotated Dataset With PCL Towards Vulnerable Communities

3.1 Introduction

In this chapter, we introduce the Don't Patronize Me! (DPM!) dataset, which is aimed at supporting the development of NLP models for identifying and categorizing language that is patronizing or condescending towards vulnerable communities (e.g. refugees, homeless people or poor families). The DPM! dataset constitutes the seed data for our research in the framework of this thesis and was released in its first version in December 2020 [125]. In 2022, we published a dedicated test set, which complements the data presented in this chapter and which will be introduced in Chapter 6.

In the next sections, we explain the process of creating the DPM! dataset, including motivation, task definition, data curation and annotation. We analyze the potential biases that we might have introduced during these processes, and expose the perceived limitations of this contribution.

3.2 Motivation and Task Definition

The motivation behind the construction of the DPM![1] dataset is to provide the NLP community with a novel resource to encourage research on a new task, namely the detection and classification of PCL towards vulnerable communities in the media.

In the previous chapter we defined PCL as a kind of language that shows a superior attitude towards others, or treats them with compassion or pity. In the context of vulnerable communities and the media, PCL is often used unconsciously and with the intention to help, for instance by raising awareness or funds, or moving the audience to action. Interpreting a message as condescending might also depend on the reader and their background. These are some of the reasons why PCL detection poses a challenge both for NLP systems and human annotators. With the objective of clarifying the purpose of the task, we next include some hints to correctly detect such kind of language.

3.2.1 How to identify PCL?

In this work, we analyze discourse on vulnerable communities. We will consider a piece of text as containing PCL when, referring to an underprivileged individual or community, we can identify one or several of the following traits:

- The use of the language states the differences between the *'us'* and the *'them'*. The vulnerable community is depicted as different to *us*, with other experiences and life stories. This discourse establishes an invisible distance between the two communities.
- The language raises a feeling of pity towards the vulnerable community, for example by abusing adjectives or by recurring to flowery words to depict a

[1] Available under request for research purposes at `https://github.com/Perez-AlmendrosC/dontpatronizeme`.

certain situation in a literary way (e.g., metaphors, euphemisms or hyperboles).

- The author and the community they belong to are presented as *saviours* of those in need. Not only do the first have the capacity to solve the problems of the vulnerable, but also a moral responsibility to do so. The superior or privileged community is also presented as having the knowledge and experience to face and solve the problems of the disadvantaged.

- In the opposite direction, the members of the vulnerable community are described as lacking the privileges the author's community enjoys, or even the knowledge or experience to overcome their own problems. They will need, therefore, the help of others to improve their situation.

- The vulnerable community and its members are presented either as victims (i.e. overwhelmed, victimized or pitied) or as heroes just because of the situation they face.

3.2.2 What is not PCL?

It can be easy to classify a piece of text as condescending towards vulnerable communities mistakenly. We want to highlight, in particular, the following two situations where the language that is used to talk about underprivileged groups is not condescending.

- Because they might be experiencing vulnerability, the news about underrepresented groups often depict rough situations. The description of an extreme situation can be harsh and stark and leave the reader with a feeling of sadness and helplessness, while not necessarily being condescending.

- With PCL, the superiority of the author is concealed behind a friendly or compassionate approach towards the situation of vulnerable communities. Thus, a

message which is openly offensive, aggressive or containing prejudiced, discriminatory or hate speech is not considered to be PCL for the purpose of our dataset.

3.2.3 Categories of PCL towards vulnerable communities

Inspired by the extensive research on PCL discussed in Chapter 2, we propose a novel taxonomy of linguistic techniques to express condescension towards vulnerable communities. The taxonomy includes seven PCL categories, which derive from three higher-level categories.

The saviour. The community to which the author and the majority of the audience belong is presented in some way as *saviours* of those vulnerable or in need. The language used subtly positions the author in a better, more privileged situation than the vulnerable community. They express the will to help them, from their superior and advantageous position. There is a clear difference between the *we* and the *they*. As part of *the saviour*, we can find examples of the following categories:

- **Unbalanced power relations.** By means of the language, the author distances themselves from the community or the situation they are talking about, and expresses the will, capacity or responsibility to help them. It is also present when the author entitles themselves to give something positive to others in a more vulnerable situation, especially when what the author *concedes* is a right which they do not have any authority to decide to give. The next sentences are examples of unbalanced power relations:

  'You can make a difference in their lives' or *'They come back in with nothing and we need to outfit them again'* or *'They deserve another opportunity'* or *'They also have the right to love'*.

- **Shallow solution.** A simple and superficial charitable action by the privileged community is presented either as life-saving/life-changing for the underprivileged one, or as a solution for a deep-rooted problem. These are some examples of shallow solutions:

  > *'Raise money to combat homelessness by curling up in sleeping bags for one night'* or *'If every supporter on Facebook donated just one box each it would make a real difference to many poor families'*.

The expert. The underlying message is that the privileged community, which the author and their audience belong to, knows better what the vulnerable community needs, how they are or what they should do to overcome their situation. We consider the following categories:

- **Presupposition**, when the author assumes a situation as certain without having all the information, or generalises their or somebody else's experience as a categorical truth without presenting a valid, trustworthy source for it (e.g. a research work or survey). Stereotypes or clichés are also considered to be examples of presupposition. As examples of presuppositions, consider the following:

  > *'[...] elderly or disabled people who are simply unable to evacuate due to physical limitations'* or *'If the economy fills with women, it will develop beautifully'*.

- **Authority voice**, when the author stands themselves as a spokesperson of the group, or explains or advises the members of a community about the community itself or a specific situation they are living. This category can be found in the following sentences:

'Accepting their situation is the first step to having a normal life' or *'We also know that they can benefit by receiving counseling from someone who can help them understand'*.

The poet. The focus is not on the *we* (author and audience), but on the *they* (the individual or community being referred to). The author uses a literary style to describe people or situations. They might, for example, abuse adjectives or rhetorical devices to either present a difficult situation as somehow beautiful, something to admire and learn from, or they might carefully detail its roughness to touch the heart of their audience. The categories we establish for *the poet* are:

- **Metaphors** can conceal PCL, as they cast an idea in another light, making a comparison between unrelated concepts, often with the objective of depicting a certain situation in a softer way. For the annotation of this dataset, euphemisms are considered to be an example of metaphors. Some examples of metaphors are:

  'Poor children might find more obstacles in their race to a worthy future' or *'those who cling to boats to reach a shore of survival'*.

- **Compassion**. The author presents the vulnerable individual or community as needy, raising a feeling of pity and compassion from the audience towards them. It is commonly characterized by the use of flowery wording that does not aim to provide information, but to embellish an almost poetic description of vulnerability. Some examples of this category are as follows:

  'Some are lured by corrupt "agents", smuggled across the searing Sahara and discarded in the streets of Europe, resigned to selling fake designer bags as undocumented immigrants' or *'For the roughly 2,000 migrants who call it home, the broken windows and decaying*

walls of the decrepit warehouse offer scant respite from the harsh blizzard conditions currently striking Serbia'.

- **The poorer, the merrier**. The text is focused on the community, especially on how the vulnerability makes them better (e.g. stronger, happier or more resilient) or how they share a positive attribute just for being part of a vulnerable community. The message expresses the idea of vulnerability as something beautiful or poetic, and people living vulnerable situations as having values to admire and learn from. We can think of the typical example of 'poor people are happier because they don't have material goods'. The next sentences contain expressions of *the poorer, the merrier*:

  'He is reminded of the true meaning of hope by people living in situations the world would see as hopeless' or *'her mom is disabled and living with her gives her strength to face everyday's life'* or *'refugees are wonderful people'*.

Figure 3.1 summarizes our taxonomy of PCL categories.

3.3 Data curation

The Don't Patronize Me! dataset contains 14,299 paragraphs about potentially vulnerable groups, annotated to indicate the type of PCL contained in them, if any. The corpus is divided into 10,467 paragraphs which are used for training and 3,832 paragraphs which are reserved for testing.

The paragraphs of the DPM! dataset have been extracted from news stories from the News on Web (NoW) corpus [34][2]. This original corpus contains more than 18 million articles crawled from online media sources in 20 English-speaking countries from 2010 until 2018. Table 3.1 shows the countries covered by our dataset.

[2]The corpus is used with the permission of its author.

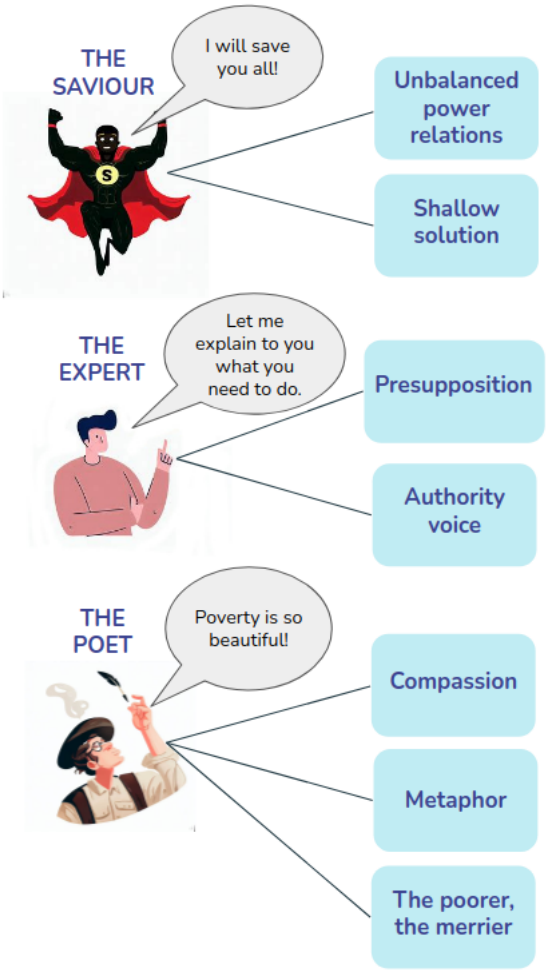

Figure 3.1: Taxonomy of PCL categories

In order to extract the paragraphs for our dataset, we first selected ten keywords related to potentially vulnerable communities widely covered by the media and susceptible of receiving a condescending or patronizing treatment. Table 3.2 shows the communities or keywords covered in this dataset.

Next, we retrieved paragraphs in which these keywords are mentioned, choosing a similar number of paragraphs for each of the 10 keywords and each of the 20 English-speaking countries that are covered in the corpus. An overview of the dataset, including the number of paragraphs for each keyword-country combination, as well as other statistical information, can be found in Chapter 4. For the Don't Pat-

Countries			
Australia	India	New Zealand	South Africa
Bangladesh	Ireland	Nigeria	Sri Lanka
Canada	Jamaica	Pakistan	Tanzania
Ghana	Kenya	Philippines	UK
Hong Kong	Malaysia	Singapore	United States

Table 3.1: Countries represented in the Don't Patronize Me! dataset. All articles included in the dataset are written in English.

Communities	
Disabled	In need
Homeless	Poor families
Hopeless	Refugees
Immigrant	Vulnerable
Migrant	Women

Table 3.2: Communities or keywords represented in the Don't Patronize Me! dataset

ronize Me! dataset, we randomly select the dates and media from the available articles in the original corpus, as we focus just on the presence of our keywords or communities. [3]

The data was annotated by three expert annotators, with backgrounds in communication, media and data science. Two annotators annotated the whole dataset (*ann1* and *ann2*), while the third one (*ann3*) acted as a referee to provide a final label in case of strong disagreement. In the next section, we explain the annotation process and challenges.

3.4 Annotation

The annotation process for this dataset posed, not unexpectedly, a significant number of challenges. PCL is a very subtle and subjective language, whose interpret-

[3]We purposely do not include this meta information in our dataset, as we want to avoid the public tracing back the articles to their original media, as we do not pursue the objective of pointing to any specific media as being especially condescending in its treatment to vulnerable communities or towards a specific group. This information, however, could be retrieved from the original corpus.

ation might depend on the receiver's (or annotator's) profile. For this reason, we carefully selected and trained our annotators[4] through a series of interviews, meetings, annotation guidelines and conflict resolution cycles. These served, in turn, to adapt the annotation guidelines so they could cover the new issues and doubts which arose in the earliest stages of the annotation. This process of training, evaluation and adaptation pursued two main objectives: 1) obtaining the highest possible common understanding of PCL and the annotation task, but also 2) preserving the individual perception of each annotator regarding PCL, so the annotation would not be, as much as possible, influenced by the research design. From an early stage, we discarded the potential use of crowdsourcing for the annotation of this dataset, due to the intrinsic challenges of the task. The difficulty and subjectivity of the annotation require a thorough training and conflict resolution process, as well as knowing and trusting the annotators' profiles and backgrounds for transparency reasons. This poses additional challenges which also hinders the potential scaling of the annotation.

To annotate the dataset, a two-step process was followed. In the first step, annotators determined which paragraphs contain PCL. Subsequently, in the second step, the annotators indicated which text spans within these paragraphs contained the condescending message, and labelled each of these text spans with a particular PCL category from the taxonomy presented in Section 3.2.3. We now discuss these two steps in more detail.

3.4.1 Step 1: Paragraph-Level Identification of PCL

The aim of this annotation step is to decide for each paragraph whether or not it contains PCL. This annotation step proved more difficult than expected, stemming from the often subtle and subjective nature of PCL. To mitigate this, we decided to annot-

[4]The annotators were postgraduate students, who were recruited via Cardiff University's Job Shop.

ate the paragraphs with three possible labels: 0, meaning that the paragraph does not contain PCL, 1, meaning that it is considered to be a borderline case, or 2, meaning that it clearly contains PCL. We computed the Kappa Inter-Annotator Agreement (IAA) between the two main annotators (*ann1* and *ann2*) across the three labels, obtaining a moderate agreement of 41%. However, if we omit all paragraphs which were marked as borderline by at least one annotator, the IAA reaches a substantial 61% [94].

To maximize the amount of information captured by the annotations, and in particular to obtain a finer-grained assessment of borderline cases, we combined the labels provided by the two annotators into a 5-point scale, as follows:

- Label 0: both annotators assigned the label 0 (0 + 0).
- Label 1: one annotator assigned the label 0 and the other assigned the label 1 (0 + 1).
- Label 2: both annotators assigned the label 1 (1 + 1).
- Label 3: one annotator assigned the label 2 and the other assigned the label 1 (2 + 1).
- Label 4: both annotators assigned the label 2 (2 + 2).

Note how partial disagreement between the annotators is thus reflected in the final label. The cases of total disagreement, where one annotator labeled the instance as clearly not containing PCL and the other annotated it as clearly containing PCL (0 + 2), were annotated by *ann3*. After this supplementary annotation, the paragraph is either labelled as 1, if the third annotator considered the paragraph not to contain PCL, as 2, if they considered it to be a borderline case, or as 3, if they considered the paragraph to clearly contain PCL. In this way, the labels 0 and 4 remain reserved for clear-cut cases.

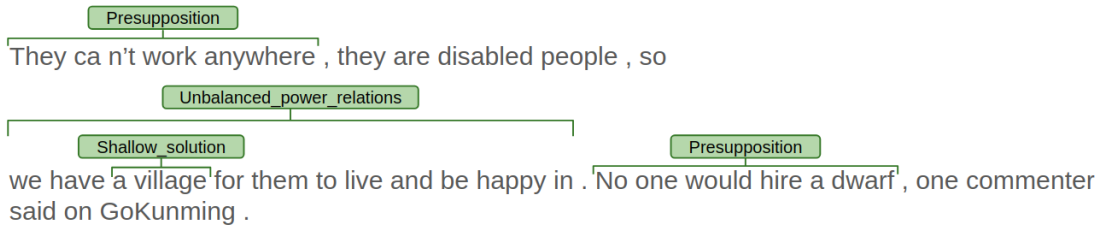

Figure 3.2: Example of categories annotation using the BRAT rapid annotation tool

3.4.2 Step 2: Identifying Span-Level PCL Categories

Those paragraphs labelled as containing PCL in Step 1 are collected for further annotation. The aim of this second step is to specify which text spans within these paragraphs contain PCL and to identify which PCL categories these text spans belong to. For this step, we used the BRAT rapid annotation tool [161][5]. An example of this annotation tool is shown in Figure 3.2. Note that each paragraph might contain one or more text spans with PCL, which may be assigned to the same or to different categories.

In addition to the seven categories presented in section 3.2.3, we also included an "Other" category for annotation purposes, to classify all the text spans which the annotators considered to contain PCL, but which they could not assign to any of the previous categories. However, no instance was labelled with "Other".

For this second step of the annotation, we compute the IAA for each category, reaching the following agreements: *Unbalanced power relations*: 58.43%; *Authority voice*: 48.34%; *Shallow solution*: 56.50%; *Presupposition*: 52.94%; *Compassion*: 66.40%; *Metaphor*: 52.72%; and *The poorer, the merrier*: 66.72%. When computing the agreement for the three higher-level categories, we obtain an IAA of 63.02% for *The Saviour* (*Unbalanced power relations* and *Shallow solution*), 57.21% for *The Expert* (*Presupposition* and *Authority voice*), and 66.99% for *The Poet* (*Compassion*,

[5]https://brat.nlplab.org/

Metaphor and *The poorer, the merrier*).

3.5 Ethical Considerations and Limitations

As we mentioned in the introduction of this chapter, the DPM! dataset aims to provide the community with a resource to support research on unfair communication towards vulnerable communities. Although we consider that this dataset might be useful for this purpose, we also recognize that it presents limitations, and there are ethical considerations that should be taken into account.

As with any other dataset, biases are unavoidably introduced in the data. In the case of the DPM! dataset, the biases and limitations start in the research design and data collection process, as we have selected only 10 keywords to show the use of PCL by the media. However, PCL is used with more communities or individuals and the language used to refer to other social groups might be different to the one represented in this dataset. In addition, our approach to extracting paragraphs about the aforementioned communities is very naïve, as these groups can be referred to through different lexical choices. The representation of media sources and countries is also limited to a small sample, and all the data collected is written in English. To mitigate this, we have tried to represent different geographical origins and cultures by covering news from twenty countries and by randomly selecting the sources of the paragraphs, without considering the media publisher. Moreover, the annotation process also presents biases and limitations. For instance, all three annotators participating in the annotation of the training data are European females, with ages between 25 and 35 years old and postgraduate studies. For the annotation of the test set, which will be presented in Chapter 6, two male annotators also participated in the process, with similar demographics. All annotators are proficient in English, although English is not the first language for any of them. An extended data statement [10] about the corpus was published together with the dataset and

can be found at https://github.com/Perez-AlmendrosC/dontpatronizeme.

3.6 Summary

In this chapter, we have presented the Don't Patronize Me! dataset, covering the motivation behind its creation and the processes of data selection, extraction and annotation. We have also included a more thorough explanation of the task of detecting and categorizing PCL and presented our novel taxonomy to classify PCL techniques.

With the objective of working towards more transparent and responsible research, we have also exposed the biases we might have introduced in the data, and the limitations and ethical considerations we have detected on it, so they can be taken into account when used by other researchers.

In the next chapter, we will present the DPM! dataset in more detail, with quantitative and qualitative analysis of the data. We will also introduce some baseline models which are trained on DPM! to detect and categorize PCL.

Chapter 4

Qualitative and Quantitative Analysis of the Data

4.1 Introduction

After introducing the DPM! dataset in Chapter 3, this chapter aims at obtaining a more in-depth understanding of the corpus, as well as analyzing its possibilities as a resource for NLP models. In Section 4.2, we offer a quantitative and qualitative analysis of the dataset, exploring the distribution of the data and its annotations. We look at the presence or absence of PCL for each community, as well as the prevalence of the different PCL categories for each vulnerable group. After getting some insights from the data, we present baseline experiments for PCL classification and categorization in Section 4.3. Specifically, we explain our experimental setting and discuss the quantitative results of a number of NLP models in Section 4.3.1. Afterwards, we perform a qualitative analysis of some of these results in Section 4.3.2.

4.2 The DPM! Dataset at a Glance

In this section, we offer an overview of the DPM! dataset. Note that these statistics correspond to the training set of the corpus, as the test set will be presented in

	dis	hom	hop	imm	mig	need	fam	ref	vul	wom	**Total**
Australia	47	51	52	56	57	56	53	54	60	55	541
Bangladesh	50	55	45	49	56	51	46	52	55	53	512
Canada	51	53	52	51	47	52	55	56	61	52	530
Ghana	62	55	57	53	58	51	25	52	54	55	522
Hong Kong	59	53	32	50	57	55	22	49	52	61	490
India	30	52	62	58	52	57	52	58	59	50	530
Ireland	61	50	55	58	58	58	36	58	48	55	537
Jamaica	53	62	47	54	50	58	11	54	50	51	490
Kenya	52	51	55	55	53	50	55	49	57	61	538
Malaysia	58	48	46	53	57	62	53	58	60	51	546
New Zealand	62	45	61	51	56	48	50	49	49	47	518
Nigeria	52	60	49	52	55	52	47	56	59	55	537
Pakistan	50	55	51	51	57	58	57	56	54	56	545
Philippines	61	56	56	48	59	54	53	51	55	52	545
Singapore	51	56	52	56	58	59	54	45	54	50	535
South Africa	60	54	57	58	54	54	59	50	47	56	549
Sri Lanka	52	57	57	51	52	48	32	56	49	50	504
Tanzania	9	54	18	51	45	49	38	48	52	51	415
UK	55	50	47	55	53	56	58	57	58	51	540
United States	53	60	54	51	54	54	53	59	47	58	543
Total	1028	1077	1005	1061	1088	1082	909	1067	1080	1070	10467

Table 4.1: Number of paragraphs per keyword and country in the dataset.

Chapter 6. The training set of the DPM! dataset, as stated in Chapter 3, contains 10,467 paragraphs in English, extracted from media sources in twenty countries. These paragraphs cover news stories about ten specific communities or keywords potentially related to vulnerable situations. Table 4.1 shows the number of paragraphs for each country-community combination[1], as well as the total number of paragraphs by community. Originally, we aimed at retrieving the same number of paragraphs for each community and country, however we encountered that some keywords were hardly mentioned in some countries, as is the case of *poor families* or *hopeless*. This might derive from one of the limitations of our research design, as the community is probably covered by the media, but it might be referred to with other lexical choices. For other keywords, such as *disabled*, some paragraphs initially extracted from the NoW corpus were excluded after a pre-processing phase, as they referred to other topics or communities, for instance, the list of disabled players

[1]The considered keywords are disabled (dis), homeless (hom), hopeless (hop), immigrant (imm), migrant (mig), in-need (need), poor-families (fam), refugees (ref), vulnerable (vul) and women (wom).

	dis	**hom**	**hop**	**imm**	**mig**	**need**	**fam**	**ref**	**vul**	**wom**	**total**
Pos. samples	81	178	124	30	176	36	150	86	80	52	993
% Label 2	23,5	14,6	6,5	16,7	10,8	13,9	18,0	18,6	8,8	23,1	14,5
% Label 3	45,7	42,1	50,0	60,0	43,2	44,4	48,0	40,7	51,3	50,0	46,1
% Label 4	30,9	43,3	43,5	23,3	46,0	41,7	34,0	40,7	40,0	26,9	39,4

Table 4.2: Number of paragraphs containing PCL per category. We also present the percentage of paragraphs annotated with 2, 3 and 4.

for a specific sport's game.

From the resultant 10,467 paragraphs contained in the train set, 993 were annotated as expressing condescension to some extent (i.e., they obtained a final label of 2, 3 or 4 after the annotation process). Table 4.2 shows how these paragraphs containing PCL are distributed among communities, as well as the robustness of the final label, represented by the percentages of labels 2, 3 and 4 for each group.[2] From this analysis, we observe that borderline cases (i.e., both annotators annotated the paragraph with label 1) are considerably less frequent than those cases where at least one of the annotators found a strong example of PCL (labels 3 or 4). However, including a borderline label proved useful, as can be seen in the percentage of paragraphs which obtained a final label of 3 and which generally surpass those with a rotund annotation (2-2). By looking at the different communities, we see how *disabled* and *women* present more borderline cases, which might reflect more subtle PCL towards these groups. On the other side, *migrants*, *hopeless* and *homeless* seem to present more flagrant cases of condescension. The high percentages of paragraphs annotated with 3 support the idea of the subjectivity of PCL and, thus, the difficulty of the task.

These 993 positive cases of condescension were in turn fleshed out in 2,760 spans expressing PCL with a specific category. In Table 4.3 we show how many spans have been labelled with each of the categories for each community, as well as the per-

[2]Note that we purposely do not show the number of positive examples or categories per country, as the objective of this thesis is analyzing PCL towards vulnerable communities, not pointing out to any region for their use of PCL.

	unb	shal	pre	auth	met	comp	merr	Total
Dis	70 (35,9%)	15 (7,7%)	23 (11,8%)	21 (10,8%)	14 (7,2%)	42 (21,5%)	10 (5,1%)	195
Hom	175 (38.3%)	61 (13.3%)	29 (6.3%)	27 (5.9%)	44 (9.6%)	117 (25.6%)	4 (0.9%)	457
Hop	77 (17.7%)	5 (1.1%)	83 (19.1%)	47 (10.8%)	46 (10.6%)	172 (39.5%)	5 (1.1%)	435
Imm	21 (25.6%)	3 (3.7%)	18 (22%)	7 (8.5%)	4 (4.9%)	25 (30.5%)	4 (4.9%)	82
Mig	29 (29%)	4 (4%)	8 (8%)	13 (13%)	9 (9%)	33 (33%)	4 (4%)	100
Need	258 (54.5%)	67 (14.2%)	14 (3%)	37 (7.8%)	29 (6.1%)	64 (13.5%)	4 (0.8%)	473
Fam	142 (32.3%)	32 (7.3%)	55 (12.5%)	51 (11.6%)	52 (11.8%)	99 (22.6%)	8 (1.8%)	439
Ref	71 (33.6%)	23 (10.9%)	17 (8.1%)	16 (7.6%)	17 (8.1%)	63 (29.9%)	4 (1.9%)	211
Vul	88 (39.3%)	9 (4%)	19 (8.5%)	39 (17.4%)	25 (11.2%)	43 (19.2%)	1 (0.4%)	224
Wom	37 (25.7%)	8 (5.6%)	30 (20.8%)	27 (18.8%)	10 (6.9%)	24 (16.7%)	8 (5.6%)	144
Total	968 (35.1%)	227 (8.2%)	296 (10.7%)	285 (10.3%)	250 (9.1%)	682 (24.7%)	52 (1.9%)	2760

Table 4.3: Number and % of text spans that have been labelled with each of the PCL categories, per keyword. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

centages over the total number of PCL spans for that community. Note how *in-need* presents the highest number of annotated spans, despite being a community with few positive examples of PCL. *Homeless*, *poor-families* and *hopeless* also present a significant number of PCL messages. Regarding the different categories, this overview shows, on the one hand, how *Unbalanced Power Relations* and *Compassion* are the most commonly used strategies to express condescension, with a 35.1% and a 24.7% of the total number of annotated spans, respectively. On the other hand, the dataset only contains 52 examples of *The poorer, the merrier*, which suppose only the 1.9% of the samples. Considering the use of categories in relation to the different communities, we can also see how *Unbalanced power relations* are especially expressed towards people *in need* and *The poorer, the merrier*, although scarce in the data, is especially used towards *women* and *disabled*. Media seems to use *Presuppositions* especially with *immigrants* and *women*, while *hopeless* people, *immigrants* and *migrants* tend to receive a tone of *Compassion* from journalists.

4.3 Classifying and Categorizing PCL: Baselines

In this section, we present the results of some baseline experiments to assess the potential of the dataset as a resource for research in modelling PCL. We experiment with a number of different methods in two settings: predicting the presence of PCL, viewed as a binary classification task (Task 1), and predicting PCL categories, viewed as a multi-label classification task (Task 2). After comparing the results of the different approaches, we analyze some examples in a qualitative way to review the performance of one of the best models.

4.3.1 Experiments and results

For our baseline experiments, we evaluate the following methods:

- **SVM-BoW.** We use a TF-IDF weighted Bag-of-Words representation of the paragraphs as input to a Support Vector Machine (SVM) implemented with SciKit-Learn [123]. For Task 1, the parameters that were selected after hyper-parameter tuning were C=10, gamma= 'scale', kernel= 'rbf', while for Task 2 we found that C=100, gamme='scale', kernel= 'linear' yielded the best results on the validation data.

- **SVM-WV.** We use paragraph embeddings as the input for a SVM, also implemented with SciKit-Learn [123]. To create the paragraph embeddings, we use the average of the standard 300-dimensional Word2Vec Skip-gram word embeddings trained on the Google News corpus [107]. In this case, the hyper-parameters that were selected are C=10, gamma='scale', kernel='poly' for Task 1 and C=100, gamma='scale', and kernel='rbf' for Task 2.

- **BiLSTM.** We use a bidirectional LSTM, using the same word embeddings as in the SVM experiment to represent the individual words. As hyper-parameters,

we use 20 units for each LSTM layer and a dropout rate of 0.25% at both the LSTM and classification layers. We train for 300 epochs, using the Adam optimizer, with early stopping and a patience of 10 epochs.

- **Fine-tuned Language Models.** We fine-tune a number of pre-trained language models for sequence classification. Specifically, we explore BERT-base and BERT-large [38], RoBERTa-base and RoBERTa-large [98], which can be viewed as an optimized version of BERT, and a DistilBERT [149] model, which is a lighter and faster variant of BERT. In all cases, we train the model for 5 epochs with a batch size of 4 and report the average of 5 runs, with the same fixed random seeds for all models.

- **Random.** To put the results in context, we include a classifier that relies on random guessing, choosing the positive class with 50% probability in Task 1, and independently selecting each label with a probability of 50% in Task 2.

For both Task 1 and Task 2 we used 5-fold cross-validation for all the experiments. For the BiLSTM models, we used 10% of the training data in each fold as a validation set for early stopping. For the SVM models, we instead tuned the hyper-parameters using Grid Search Cross-Validation[3]. As mentioned before, for Task 1 we view paragraphs labelled with 0 or 1 as negative examples, and the remaining paragraphs, labelled with 2, 3 or 4, as positive examples. The results are reported in terms of the precision, recall and F1 score of the positive class. Task 2 is viewed as a paragraph-level multi-label classification problem, where each paragraph is assigned a subset of the PCL category labels. Therefore, in these baselines, span boundaries are not used as part of the training data. We report the precision, recall and F1 score of each of the individual category labels.

The results of Task 1 are summarized in Table 4.4. As can be seen, all of the considered methods clearly outperform the random baseline. Unsurprisingly, Language

[3]We used the GridSearchCV module of the Scikit-Learn library to select the best hyper-parameters in a 5-fold cross-validation setting.

	P	R	F1
Random	10.08	52.67	16.92
SVM-BoW	41.61	33.74	37.25
SVM-WV	35.76	51.67	42.26
BiLSTM	35.32	49.99	40.48
DistilBERT	47.16	60.64	53.03
BERT-base	46.94	64.43	54.28
BERT-large	**49.10**	63.17	**55.11**
RoBERTa-base	45.41	**68.34**	54.48
RoBERTa-large	43.83	60.34	50.73

Table 4.4: Results for the problem of detecting PCL, viewed as a binary classification problem (Task 1).

Models achieve the best results, with BERT-large obtaining the best results and performing slightly better than RoBERTa-base. RoBERTa-large obtains very unstable results, as some folds of the cross-validation setting would outperform any other approach, while in others the model does not seem to learn anything. Table 4.5 shows the results obtained in Task 2. RoBERTa-large outperforms the rest of the models in all the categories except for *Compassion*, where BERT-base gets the best results. We can also notice the general poor performance of the BiLSTM, with SVMs models obtaining better results across many categories. The SVM-WV model stands out for some categories, such as *Metaphors*, where it outperforms DistilBERT, BERT-base, BERT-large and the BiLSTM results. For *The poorer, the merrier* this approach outperforms all the other models except for RoBERTa-large. These results seem to point to SVM-WV performing especially well for labels with few examples.

Comparing the results for different categories, we can see that *Unbalanced power relations* seem relatively easy to detect. This is not unexpected, given that the presence of words such as *us*, *they*, *must* or *help* are strong and common indicators of such language. For similar reasons, instances of *Compassion* appear relatively easy to detect. *The poorer, the merrier* is the least represented category in the entire dataset, with just 52 samples, which can explain the poor results for this category. However, the poor performance for the *Metaphor* category by most models cannot be explained in this way, given that the number of training examples for this category

	Random			SVM-BoW			SVM-WV		
	P	R	F1	P	R	F1	P	R	F1
Unb	73.19	52.23	60.96	79.34	78.38	78.83	81.7	84.1	82.9
Shal	18.46	48.98	26.82	40.95	37.39	38.86	59.6	46.6	52.1
Pres	22.47	49.55	30.92	43.56	39.97	41.06	49.1	43.0	45.4
Auth	23.29	51.74	32.12	36.15	36.40	35.88	41.5	37.8	39.4
Met	20.25	49.24	28.7	30.79	30.2	30.32	43.7	33.1	37.3
Comp	46.94	50.75	48.77	64.32	60.94	62.50	72.6	70.8	71.6
Merr	4.44	55	8.22	4.00	2.86	3.33	20.0	10.1	13.0
	BiLSTM			**DistilBERT**			**BERT-base**		
	P	R	F1	P	R	F1	P	R	F1
Unb	82.96	86.13	84.36	85.35	89.76	87.46	84.54	92.43	88.30
Shal	60.91	37.03	45.06	70.29	50.86	58.91	74.65	52.96	61.77
Pres	53.88	30.14	36.98	62.70	45.97	52.69	61.77	54.46	57.09
Auth	41.48	15.27	21.50	56.17	34.82	42.57	56.07	41.91	47.71
Met	27.16	4.12	6.65	54.27	15.40	23.40	59.60	25.47	35.36
Comp	73.83	68.97	71.01	80.09	72.03	75.79	81.21	75.26	**78.06**
Merr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BERT-large			**RoBERTa-base**			**RoBERTa-large**		
	P	R	F1	P	R	F1	P	R	F1
Unb	85.22	89.76	87.38	86.47	90.91	88.59	87.13	92.65	**89.79**
Shal.	72.23	49.58	58.64	70.29	51.01	58.89	72.07	54.53	**61.93**
Pres	62.02	45.28	52.08	62.03	49.57	54.30	65.23	55.57	**59.42**
Auth	59.26	39.57	47.35	61.32	47.27	53.08	60.66	48.56	**53.58**
Met	58.80	10.75	18.03	53.89	30.69	38.56	54.18	36.17	**43.16**
Comp	80.82	72.32	76.24	79.06	74.68	76.78	79.89	75.22	77.43
Merr	0.00	0.00	0.00	0.00	0.00	0.00	50.40	16.58	**23.95**

Table 4.5: Results for the problem of categorizing PCL, viewed as a paragraph-level multi-label classification problem (Task 2). The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

is higher than the number of examples for *Shallow solution* and very similar to the number of examples for *Authority voice*. More generally, while some of the differences in performance are due to variations in the number of training examples, the categories with the weakest performance also tend to be those that require some form of world knowledge. For instance, to detect presuppositions or the use of authoritarian voices, we need to determine whether the assumption which is made is reasonable or not, and probably understand the power relations and inequalities involved. Similarly, detecting shallow solutions requires assessing the quality of the proposed solution, which can clearly be challenging. Surprisingly, *The poorer, the merrier* was the category which showed the highest IAA among the annotators,

so although posing clear challenges to NLP models, it was easily identified by humans. Similarly, the models struggle to understand what makes a solution shallow and condescending, while humans generally agree in pointing out these instances. *Presuppositions*, *Metaphors* and, especially, *Authority Voice* posed a challenge both for automatic models and humans.[4]

The main takeaway from this analysis is that, among automatic approaches, Language Models allow us to obtain the best results in the tasks of classifying and categorizing PCL, with their large versions, namely BERT-large and RoBERTa-large, outperforming the other approaches in Task 1 and Task 2 respectively. However, large models present some limitations, such as the computational cost of the experiments and the occasional inconsistency in their performance, as we have seen specifically with RoBERTa-large. Thus, throughout this thesis, we will consider a smaller and more robust model, namely RoBERTa-base, as the baseline for further research and analysis.

4.3.2 Qualitative analysis

To get further insights into the dataset and the performance of our baseline model, we collect and analyse specific examples of the predictions of RoBERTa-base. Table 4.6 shows some examples of paragraphs from Task 1, their gold labels and the predicted label by our baseline model. There are three correctly classified instances and seven misclassified examples (i.e. four false positives and three false negatives). The three first paragraphs, correctly classified as containing PCL, present clear-cut examples of condescension towards vulnerable communities, with expressions of *The saviour* (e.g. "McDonald's helps feed homeless"), *The expert* (e.g. "poor parents who repeatedly make bad decisions to their children", a suggestion which comes from someone who had "real-life encounters with poor families" and feels now entitled to give advice about *them*) and *The poet* (e.g., with figurative language

[4] See Section 3.4.2

Pred.	Paragraph	Gold
pos.	After Vatican controversy, McDonald's helps feed homeless in Rome.	pos.
pos.	From his personal story and real-life encounters with poor families, manpower correspondent Toh Yong Chuan suggested shifting the focus from poor parents who repeatedly make bad decisions to their children (Lifting families Out of poverty: Focus on the children; last Thursday).	pos.
pos.	He said their efforts should not stop only at creating many graduates but also extended to students from poor Families so that they could break away from the cycle of poverty.	pos.
pos.	These shocking failures will continue to happen unless the Government tackles the heart of the problem - the chronic underfunding of social care which is piling excruciating pressure on the NHS, leaving vulnerable patients without a lifeline.	neg.
pos.	Lilly-Hue: His ability to make sure our family is never in need - his sacrificial self.	neg.
pos.	Any Kenyan small-scale farmer with such an income could not be said to be hopelessly mired in agrarian destitution. But of course, nothing in life is ever so simple as to allow for neat and precise answers.	neg.
pos.	Selective kindness: In Europe, some refugees are more equal than others.	neg.
neg.	"The biggest challenge is the no work policy. I think that refugees who come here, or asylum seekers, they're unable to work and they have kids here - their kids are stateless. That's really the cause of a lot of stress in the community."	pos.
neg.	"The people of Khyber Pakhtunkhwa are resilient. I did not see hopelessness on any face," he said.	pos.
neg.	Teach kids to give back: When Kang runs summer camps with kids, she includes "Contribution Fridays" - the kids work together as a team to make sandwiches for the homeless and dole out the food in shelters.	pos.

Table 4.6: Examples of predictions made by RoBERTa-base in Task 1.

such as "break away from the cycle of poverty"). In these cases, the model has correctly identified traits of PCL.

In the next four examples, which the model mistakenly considers as PCL, we can see words and phrases that are often used for condescension, but which are not used in a condescending context in these cases. For instance, in the fourth example, the excess of adjectives and flowery wording, e.g. *shocking failures* and *excruciating pressure*, are often used in PCL fragments from the *Compassion* category. In this example, however, it is used in a political context, without being condescending towards any particular group.

The last three examples are misclassified as not containing PCL. Consider the ninth

example, an example of the category *The poorer, the merrier*, which all models struggle to detect. Surprisingly, this category has the highest inter-annotator agreement in the annotation of the dataset. This suggests that, while for human annotators it is very easy to identify cases of this category, the models struggle to detect such cases.

In Table 4.7, some incorrect predictions from Task 2 are presented. Among others, these examples illustrate how RoBERTa-base struggles to distinguish between presuppositions and authority voices, which are often incorrectly predicted together. Shallow solutions are also often neglected by RoBERTa-base. A particularly clear case is the last example, where recognizing the presuppositions and shallow solutions in the text requires external knowledge of the situation and the needs of those affected. We can also see examples where the occurrence of a particular structure of language appears to mislead RoBERTa-base, e.g. *to open the doors wider for [...]*, in the fourth example, seems to lead the model to bet on a shallow solution. *Methaphors*, as in this same example, are also difficult to identify for RoBERTa-base in this context.

4.4 Summary

In this chapter, we have delved deeper into the Don't Patronize Me! dataset, presenting an overview of the data and the annotated labels. Also, we have presented the results of a number of baseline models for two tasks:

Task 1. Detecting whether a paragraph contains PCL, understood as a binary classification task.

Task 2. Categorizing what types of PCL are present in a paragraph expressing condescension, understood as a multi-label classification task.

Paragraph	Gold	Pred
[...] The blacks want all our farmland without compensation. Give it to them. Let the farmers flock into the cities and make a new life for themselves. With their resilience I am sure it will not be so difficult for them to establish a new, happy and productive life. They will have no money but the clothes on their back to start off with, but that is what so many immigrant Americans had to face. Through guts, determination and sheer will power, they rose above it all, and look what America is today.	unb, pres, comp, merr	unb, auth, pres, comp, met
According to the foundation, a number of children between the ages of six and 14 homeless and roaming the streets is becoming alarming.	comp	unb, comp
The photo of a Hyderabad traffic policeman feeding an elderly homeless woman has gone viral, earning him accolades from social media users [...].	unb, shal	unb
Practical ways to open the doors wider for our disabled	unb, met	unb, shal
He could have also taken his condition to mean he must be disabled from seeking to live for others. He could have degenerated into self pity as many do, wallowing in the muddy fields of self-obsession and low self esteem. Yusuf did not; everything was not about his immediate interests, but a social impact that touched even the lives of strangers [...].	unb, comp, met, merr	auth, pres, comp
She called on the general public to volunteer to donate blood and that way rescue the lives of patients in need of blood transfusion.	unb, auth	unb, auth, met
For now the families are staying with friends and family. During the day they clean up the debris left by the fire, hoping that someone will come to their rescue. They received emergency relief packs, but they are still in need of clothes, beds, blankets and kitchen appliances.	unb, shal, pres, comp	unb, comp

Table 4.7: Examples of incorrect predictions made by RoBERTa in Task 2. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

Moreover, we presented a qualitative analysis of the predictions of RoBERTa-base in order to gain a better understanding of the task.

Our exploratory analysis shows that identifying condescending or patronizing texts is a difficult challenge, both for human judges and for NLP systems. Apart from the subtle and subjective nature of PCL, a particular challenge comes from the fact that accurately modelling such language often requires knowledge of the world and common sense (e.g. to assess whether a proposed solution is shallow, or whether a particular presupposition is warranted). Nonetheless, we found that both identifying

PCL (Task 1) and categorizing occurrences of PCL (Task 2) is feasible, in the sense that non-trivial results can be achieved, with BERT-based approaches outperforming simpler methods.

In the next chapter we will explore what kind of previous knowledge would benefit a model to perform better in the tasks of identifying and categorizing PCL and what pre-training techniques are more suitable for these challenges.

Chapter 5

Pre-Training Language Models for Identifying PCL: An Analysis

5.1 Introduction

In this chapter, we explore to what extent PCL detection models can be improved by pre-training them on other, more established NLP tasks. With this approach, we aim at developing a better understanding of the nature of PCL by unveiling which types of pre-training tasks are most effective for PCL detection, if any. For instance, the hypothesis that human values are important for modelling PCL might be supported or undermined by the performance on PCL detection of a model which has been pre-trained on such data. As another example, given the subtle nature of PCL, and the fact that it is usually well-intended, it is unclear to what extent more explicit forms of harmful language, such as hate speech, can influence the modeling of PCL. To this end, we analyze the performance of PCL models pre-trained on types of language which openly try to undermine others.

In order to infuse knowledge from different tasks to PCL detection models, we consider two types of pre-training strategies, namely full fine-tuning of LM and the use of adapters [75], and apply them to ten text classification tasks. The remainder of the chapter is organized as follows: Section 5.2 offers an overview of the ten tasks on which we pre-train our model and their associated data. In Section 5.3 we intro-

duce the experimental setup of the work presented in this chapter. First, we explain the methodology followed in the experiments in Section 5.3.1 and then present the quantitative results in Section 5.3.2. In Section 5.3.3 we analyze some examples of well-classified instances by the best-performing models and try to understand how each auxiliary task improves the performance of the baseline model. To conclude the chapter, in Section 5.4 we present some of the conclusions derived from this work.

5.2 Auxiliary Datasets

Although datasets that specifically address PCL are scarce, some of its associated challenges are also addressed in other tasks. While the idea of pre-training language models on auxiliary tasks is common practice [147, 100], the success of this strategy crucially depends on the relevance of the selected tasks [135] and their relation with the target challenge, namely PCL detection in this case.

We consider four types of pre-training tasks for our experiments. First, we include tasks that involve modelling human value judgements. In particular, we consider three tasks from the ETHICS dataset. This dataset, introduced by Hendrycks et al. [73], aggregates 5 tasks involving situations that need to be classified based on human values. We focus in particular on the Commonsense Morality, Social Justice and Deontology tasks, as they follow the same format as the *Don't Patronize Me!* dataset, i.e. they are binary text classification problems. For the three selected datasets, we combine the training and test splits to train our models. However, we discard the *test hard* partition, as it contains more ambiguous instances that could confound the model. In addition to the former, we also consider the StereoSet dataset [112] which measures stereotype bias in assumptions. We now describe the aforementioned tasks in more detail:

Commonsense Morality includes 17,795 assertions about specific scenarios, which need to be classified as acceptable or not based on commonsense moral judgements.

Deontology contains 21,760 pairs of the form situation-assertion or petition-excuse, where the assertions and excuses need to be classified as being reasonable or not.

Social Justice includes 24,495 examples of the form “X deserves Y because Z", where the task is to predict whether the scenario is reasonable in terms of fairness.

StereoSet includes 6,369 instances of the form context-assumption, where the task is to predict if (i) the assumption contains stereotypes; (ii) the assumption does not contain stereotypes; or (iii) the context and assumption are unrelated.

Second, we focus on tasks that involve detecting harmful language. We focus in particular on the Hate and the Offensive datasets [7, 187], both of which are included in the TweetEval framework [5]. The details of these pre-training tasks are as follows:

Hate speech contains 27,000 tweets, which need to be classified as containing hate speech or not.

Offensive language is a collection of 14,100 tweets, where the task is to detect any kind of language that could offend either the target of the tweet or a general audience.

We also consider two datasets that focus on political language. The interest in political discourse, in this context, stems from the fact that the way in which vulnerable communities are referred to plays an important role in such discourse. Indeed, PCL has been widely studied in relation to political discourse [80]. We focus in particular

on Hyperpartisan News Detection [88] and Democrats vs Republicans Tweets[1]. The details are as follows:

Democrat vs Republican Tweets contains 86,460 tweets from US politicians, labelled as Democrat or Republican. The aim is to predict the political stance of the author of a given tweet.

Hyperpartisan News Detection is a small dataset with 645 news articles, which need to be classified as hyperpartisan or not.

Finally, as a more exploratory analysis, we also include two datasets from tasks that are intuitively less related to PCL detection, in particular the identification of irony [170] and sentiment analysis [146], both also extracted from the TweetEval framework [5]. Although the task of detecting irony may seem to have little in common with PCL detection, there are nonetheless some correspondences, such as the use of flowery and ornamented language and the prevalence of strongly opinionated inputs. Furthermore, we expect that some linguistic features that are related to the expression of sentiment might also help to detect PCL. The details of these tasks are as follows:

Irony consists of 4,601 tweets, where the task is to predict if they contain irony or not.

Sentiment consists of 59,899 tweets, where the task consists in classifying the sentiment of each input as negative, neutral or positive.

5.3 Experiments

The datasets presented in the previous section are then used to infuse knowledge into pre-trained LMs, which will then be fine-tuned on PCL. This work aims at explor-

[1] www.kaggle.com/kapastor/democratvsrepublicantweets

ing the following research questions:

1. To what extent can the performance of PCL detection models be improved by pre-training these models on auxiliary datasets?
2. Which auxiliary tasks are most effective, and what does this tell us about the nature of Patronizing and Condescending Language?
3. How does the effectiveness of the pre-training strategies vary across different PCL categories?

5.3.1 Methodology

We compare two standard strategies for pre-training a language model, namely full fine-tuning and the use of adapters [75].

Full fine-tuning. A popular approach to transfer learning when using LMs consists in fine-tuning a pre-trained model, e.g., BERT, on an auxiliary task first and then fine-tuning the resulting model on the target task. During the process of fine-tuning, the model updates its weights to adapt the representation of words to the domain of the data provided. In the case of supervised learning, the model also learns to classify the data into the provided classes. It is hoped that pre-training on an auxiliary task infuses some kind of knowledge or capability into the language model, which can then be exploited in the target task. However, an undesired consequence of pre-training in this way is the *catastrophic forgetting* of its previous knowledge that sometimes occurs [102, 142]. Figure 5.1 shows how full fine-tuning works.

Adapters. The use of adapters [75] is an alternative to full fine-tuning. In this case, new layers are added to the language model, which are trained on the auxiliary task, while the layers from the original model are frozen. Since the parameters

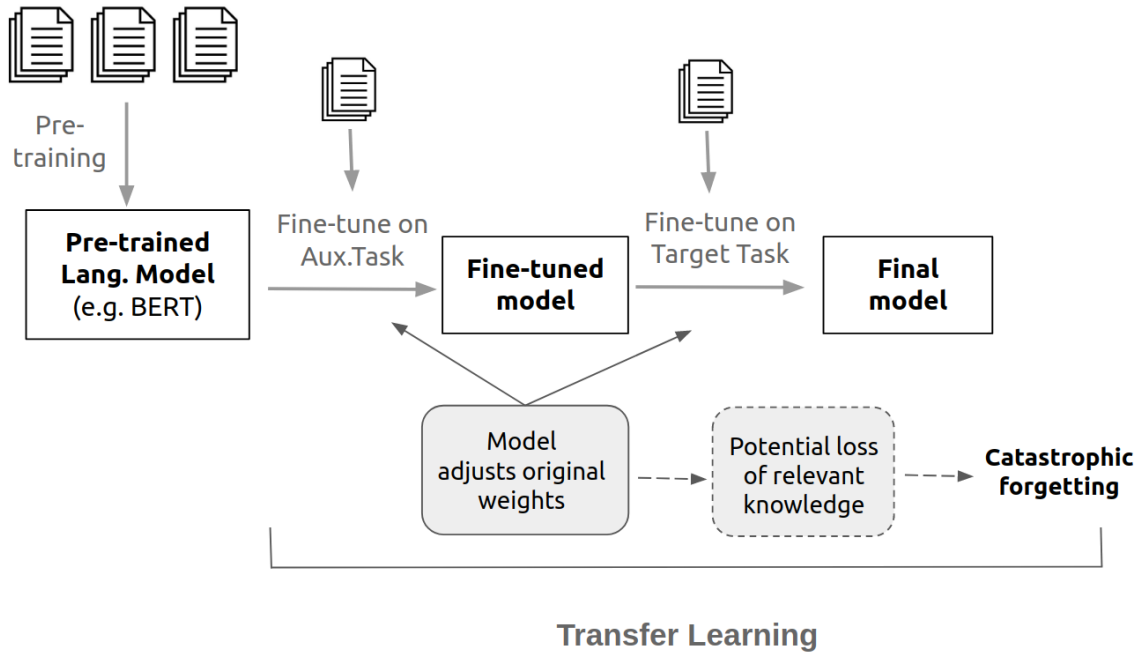

Figure 5.1: Transfer learning with full fine-tuning.

from the original language model are not updated during pre-training, catastrophic forgetting should not occur. After adding the adapter layers the model is fine-tuned on PCL detection. We consider two variants of the strategy with adapters: one in which the classification head for PCL detection is initialized based on the auxiliary task (i.e. both the adapter layers and classification head are transferred) and one in which the classification head is randomly initialised (i.e. only the adapter layers are transferred). We will refer to these variants as *Adapters+Head* and *Adapters* respectively. Figure 5.2 represents the process of transfer learning with the use of adapters.

We use the Simple Transformers library[2] for fine-tuning the models and AdapterHub [129] for training the adapters, both of which are built over the Transformers library by Wolf et al. [183].

Our experimental setting is as follows:

[2]github.com/ThilinaRajapakse/simpletransformers

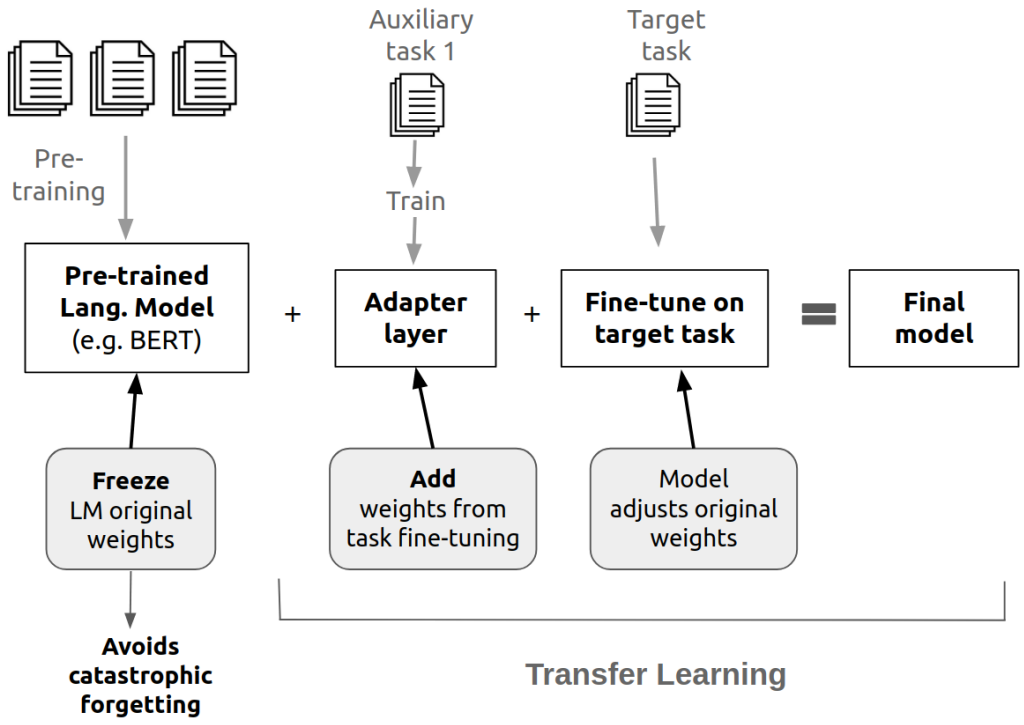

Figure 5.2: Transfer learning with the use of adapters.

Step 1: Auxiliary Task Pre-Training In each experiment, we start with a pre-trained RoBERTa-base model [98], which we train on one of the auxiliary tasks described in Section 5.2, either using full fine-tuning or using adapters. For full fine-tuning, we use a learning rate of $1e^{-5}$, following Hendrycks et al. [73]. When using adapters, we use a learning rate of $1e^{-4}$, following Pfeiffer et al. [130]. For both strategies, we use a batch size of 8 while training, which was the largest value we could fit into GPU memory. Furthermore, we fix the number of epochs depending on the size of the dataset, pre-training for 10 epochs on *Hyperpartisan* and *Irony*, as these are the smallest datasets, and for 5 epochs on the other tasks.

Step 2: PCL Fine-Tuning After pre-training on a given auxiliary task, we fine-tune the resulting model on the PCL dataset, focusing on both the binary and the multi-label classification settings, i.e. Task 1 and Task 2. As a baseline, we directly train RoBERTa-base on the PCL dataset, without infusing any previous knowledge.

As in Chapter 4, we use 5-fold cross-validation for all experiments. We train the models for 5 epochs for Task 1 and for 10 epochs for Task 2. We use a learning rate of $1e^{-5}$ and a batch size of 4.

Our main evaluation metric is the F1 score. However, to explore to what extent different categories of PCL are impacted, we also look at the recall per category, i.e. among all the paragraphs that are labelled with a particular category of PCL (e.g. UNB), we compute what percentage were correctly predicted as positive examples by the model.

5.3.2 Experimental Results

For all the experiments presented in this section, we report the average results across 5 runs, as well as the standard deviation. For each run, we do a 5-fold cross-validation setting to evaluate the model in each experiment.

Configurations that outperform the RoBERTa baseline are shown in bold.

In Table 5.1 we report the average F1 score for each of the considered auxiliary tasks, for three pre-training strategies: adapters, adapters+head and full fine-tuning. Note that for *StereoSet* and for *Sentiment*, the classification head of the adapter can not be used, as the number of labels in the auxiliary task and the main task is different.

One immediate conclusion is that using adapters outperforms fully fine-tuned models in all tasks. This suggests that catastrophic forgetting is indeed an issue in our setting, which could be related to the fact that the auxiliary tasks are only loosely related to the problem of PCL detection. In fact, *Commonsense Morality* is the only task for which full fine-tuning outperforms the baseline.

Focusing now on the strategies with adapters, in most cases, *Adapters* outperforms *Adapters+Head*. For *Adapters*, six out of ten configurations outperform the

	Adapters	**Adapters+Head**	**Fine-Tuning**
RoBERTa baseline	54.48 $_{0.42}$	54.48 $_{0.42}$	54.48 $_{0.42}$
Commonsense Morality	**55.16** $_{0.59}$	**54.53** $_{0.87}$	**54.50** $_{0.28}$
Deontology	**54.67** $_{0.56}$	54.33 $_{0.72}$	51.58 $_{0.66}$
Social Justice	53.30 $_{0.18}$	54.18 $_{0.22}$	51.72 $_{0.17}$
StereoSet	53.45 $_{0.51}$	-	53.07 $_{0.61}$
Hate Speech	**55.07** $_{0.29}$	**55.23** $_{0.31}$	54.07 $_{0.34}$
Offensive Language	**55.18** $_{0.63}$	**55.10** $_{0.44}$	52.91 $_{0.36}$
Democrat vs Republican	52.80 $_{0.80}$	53.83 $_{0.23}$	52.54 $_{0.54}$
Hyperpartisan	54.34 $_{0.41}$	53.98 $_{0.33}$	52.23 $_{0.18}$
Irony	**54.99** $_{0.40}$	**54.72** $_{0.32}$	52.56 $_{0.38}$
Sentiment	**54.97** $_{0.48}$	-	54.02 $_{0.14}$

Table 5.1: F1 score (for the positive class) on PCL Detection with different auxiliary tasks and pre-training strategies.

RoBERTa baseline, although the improvements in some cases are almost negligible. The strongest improvements are obtained for *Ofensive Language*, *Commonsense Morality* and *Hate Speech*, followed by *Irony* and *Sentiment*. In contrast, the results for *Democrat vs Republican*, *Social Justice* and *Stereoset* are weak. The results for *Irony* are surprising, as they are clearly above the baseline, despite the fact that irony detection is conceptually rather different from PCL detection. Similarly, the good results of pre-training on Sentiment are also unexpected, as we would have expected news stories to be more objective or at least contain a more subtle expression of sentiments than the tweets contained in the Sentiment dataset.

In Table 5.2, we summarize the performance of the pre-trained models for each category, considering again the average F1 score. For this table, we consider the *Adapter* strategy and fine-tune the resultant model on a multi-label classification problem, trained for 10 epochs only on the positive cases of PCL, which are in turn labelled with one or several PCL categories.

All pre-trained models improve the baseline results, especially *Offensive* and *Sentiment*, which improve the performance in five out of seven categories. *Commonsense Morality*, however, only benefits three out of seven categories, despite being one of the best-performing models in Table 5.1. In regards to specific categories,

	PCL Categories													
	UNB		**SHAL**		**PRES**		**AUTH**		**MET**		**COMP**		**MERR**	
RoBERTa base.	88.59	0.56	58.89	1.39	54.3	1.95	53.08	1.06	38.56	3.13	76.78	0.8	0.00	0.00
Comm. Mor.	**89.27**	0.31	**59.72**	1.43	52.48	0.92	52.52	1.43	34.47	1.47	**77.12**	0.93	0.00	0.00
Deontology	88.40	0.26	**61.84**	2.01	52.58	1.84	51.27	0.98	**39.97**	1.58	**78.02**	0.85	0.00	0.00
Social Justice	**88.79**	0.41	**60.90**	1.34	53.72	1.34	47.67	1.79	37.25	2.68	**77.79**	0.92	0.00	0.00
StereoSet	**89.21**	0.39	**60.11**	1.42	54.07	2.15	52.06	2.45	35.52	2.21	**77.65**	0.30	**0.89**	1.99
Hate Speech	**88.88**	0.66	**59.37**	2.29	53.84	0.65	49.34	1.66	**43.75**	2.57	**77.67**	0.29	0.00	0.00
Offensive Lang.	**89.20**	0.07	**60.60**	0.86	**55.73**	1.40	52.54	1.21	**39.04**	0.65	**78.25**	0.57	0.00	0.00
Dem. vs Rep.	**89.02**	0.21	56.95	1.09	54.01	0.93	51.33	1.02	**41.65**	2.90	**77.43**	0.79	**1.78**	2.43
Hyperpartisan	**88.74**	0.46	56.42	4.45	54.26	2.95	52.15	1.05	**40.28**	1.70	**77.42**	0.75	0.00	0.00
Irony	**88.83**	0.29	**60.28**	0.80	52.99	0.51	**53.66**	1.12	36.59	2.37	**77.68**	0.86	0.00	0.00
Sentiment	**89.23**	0.19	**59.27**	5.77	52.24	5.83	**53.55**	1.28	**39.00**	2.82	**79.04**	1.20	0.00	0.00

Table 5.2: F1 score per category for models that were pre-trained using adapters. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

COMP, *UNB* and *SHAL* see improvements after pre-training in (almost) all the auxiliary tasks, while *PRES* is only improved by the *Offensive* model, and *AUTH* only obtains better results after pre-training on *Irony* or *Sentiment*. Based on these results, a model able to identify shallow solutions in condescending language, would benefit from learning especially about deontology and social justice, which we could intuitively relate to this category, as they refer to human values, ethics and fairness; more surprisingly, offensive language and irony also provide helpful knowledge to a model to identify shallow and insufficient solutions to deep-rooted societal problems. For *MET*, we also find some interesting results, as the most important gains are obtained by the models pre-trained on *Hate Speech*, *Democrat vs Republican* and *Hyperpartisan News Detection*. All share a sometimes radicalized political discourse, which openly tries to harm or influence others. Moreover, these datasets often include insults, offences and disparagement in a more or less subtle language, which might help identify other figures of speech, such as metaphors, euphemisms or hyperboles.

However, the imbalance of samples from different categories in the training set,

	PCL Categories													
	UNB		**SHAL**		**PRES**		**AUTH**		**MET**		**COMP**		**MERR**	
RoBERTa base.	71.48	1.36	72.39	1.89	67.17	1.74	66.42	1.92	76.18	1.12	72.57	1.44	70.55	1.95
Comm. Mor.	**72.00**	1.26	**73.71**	1.35	**69.18**	2.11	**66.80**	0.76	**77.74**	1.43	71.57	1.80	68.09	3.97
Deontology	**71.57**	0.86	**74.51**	1.16	64.43	0.67	64.39	0.52	75.44	1.34	71.26	1.10	69.73	2.10
Social Justice	**73.71**	0.80	**73.33**	0.76	**68.86**	1.21	**68.54**	0.77	**76.73**	0.79	**73.47**	1.08	**72.73**	3.34
StereoSet	70.64	1.10	71.73	2.64	66.94	0.96	65.83	1.91	73.43	1.08	71.48	0.53	70.24	4.82
Hate Speech	**72.47**	0.85	**72.93**	0.81	**67.66**	0.26	**67.44**	1.26	74.67	0.66	**73.03**	1.40	**70.65**	2.91
Offensive Lang.	**71.86**	1.28	72.07	1.58	**69.04**	1.24	**66.93**	0.98	**76.86**	1.52	**73.03**	1.38	68.72	4.55
Dem. vs Rep.	69.48	1.10	69.65	1.41	63.35	0.80	64.30	1.02	75.52	0.29	70.26	1.03	**71.95**	3.87
Hyperpartisan	**71.63**	0.57	72.07	1.44	**68.50**	1.83	**67.43**	1.60	76.05	0.91	**73.45**	1.13	**71.89**	3.14
Irony	70.65	0.84	71.52	1.21	65.40	0.87	64.91	0.90	73.93	1.34	71.24	0.72	67.67	5.01
Sentiment	**71.79**	1.23	72.14	3.06	**67.43**	1.08	65.56	2.06	75.31	1.85	**73.04**	0.68	68.03	3.96

Table 5.3: Recall per category for models that were pre-trained using adapters. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

makes the multi-label classification task more difficult, especially for those categories with fewer examples, as we can see in the results for *MERR* in table 5.2. We hypothesize that some pre-trained models might be better at identifying PCL examples from a given category, even when they might not be able to identify to what category the sample belongs. In order to test this idea, we conduct a new experiment, whose results are presented in Table 5.3. For this experiment, we replicate the binary classification setting. Then, we look at which categories are present in the correctly classified instances by each model, to obtain the recall per category. As before, all results are averaged across 5 runs and with a 5-fold cross-validation setting for each run. As can be seen, the baseline is outperformed across all categories by several pre-trained models. Regarding specific categories, *MET*, *SHAL* and *MERR* seem to pose a struggle for most models, while *UNB* and *PRES* are easier to detect by most pre-trained models.

Comparing the performance of the different auxiliary tasks, we can see that, surprisingly, *Social Justice* achieves the most consistent results in recall, outperforming the baseline across all categories, in spite of its disappointing results in Table 5.1

and Table 5.2. This imbalance between the recall over categories and the overall performance on PCL detection can be explained by a low precision. With this, we could infer that pre-training on Social Justice does help to detect PCL, but also leads the model to predict false positives, which in turn leads to poor overall performance. *Hyperpartisan* presents a similar case, improving the baseline in terms of recall in five out of seven categories. By contrast, *Irony*, which presents a fairly good performance on PCL detection, does not improve the baseline for any category in terms of recall, which points to a higher precision, which would balance the overall performance. More expectedly, *Hate Speech*, *Commonsense Morality* and *Offensive Language* outperform the baseline in most categories. The worst results in Table 5.3 are for *StereoSet* and *Democrat vs Republican*, which is in line with their disappointing overall performance in Table 5.1.

5.3.3 Qualitative Analysis

To better understand how pre-training can benefit the PCL detection model, Tables 5.4, 5.5, 5.6, 5.7, 5.8 and 5.9 show some examples of paragraphs containing PCL, which were incorrectly classified by the baseline model but correctly classified by one of the pre-trained models. In the following, we try to understand what kind of knowledge from the auxiliary tasks might be helping the model to better identify PCL in each case.

The **Commonsense Morality** dataset contains various sentences about situations where the author is helping others. Some examples of such sentences are as follows:

- *I received admiration for helping an elderly woman cross the street.*
- *A troubled friend came to me for help so I listened to her.*
- *I served leftover meat to the homeless.*

Model	Text	Categories
C.Morality	“I have a lot of sympathy for folks who are in need in the city," Mr Edmonds-Waters said. “This has become an extremely expensive city to live in. The divide between those who have and those who don’t is ridiculously ginormous."	UNB, COMP
	We have done a great deal of work but the biggest thing we have done is to make the ministry a powerful tool for the protection of women... over the last four years, every time somebody comes up with a good idea or proposes an area in which women need help, we have adopted it immediately. We have been doing two things – policymaking, which also pushes other ministries, and looking at individual cases of suffering. For me, both components are equally important.	UNB, COMP
	There are also angels who get together and help a larger group of people in need.	UNB, MET
	At a ceremony held in Accra, she said the presence of disabled persons begging on the streets and the absence of ambulance to aid in the transfer of patients in need of critical help, moved her to donate the items.	UNB, SHAL, COMP
	It wasn’t uncommon for Johannah to hear through the media of an individual in need of assistance, and she would immediately make it her objective to get that person what they often desperately needed, and rarely without success.	UNB

Table 5.4: PCL paragraphs correctly classified by the model pre-trained on *C.Morality* and missed by the baseline model. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

- *I put a dollar bill in the homeless man’s empty cup.*

Comparing these sentences to the examples in Table 5.4 reveals some clear similarities. For example, in both cases, the format of the text fragments centers on personal experiences or personal actions told in first person. By pre-training on *commonsense morality*, the model thus becomes exposed to language that is characteristic of PCL, especially when it comes to recognising *UNB* (i.e. the unbalanced relation between those who give and those who receive), and *SHAL*.

The examples from the **Deontology** dataset tell us what a person in a certain situation should morally or ethically do. Positive examples from this dataset mainly talk about helping and protecting others, which is a clear trait of PCL, especially for *UNB*.

Model	Text	Categories
Deontology	“We are in a state of mourning but we have to carry on because we have a duty to those people who are so in need of our support." Mr Bird said thousands of people worldwide had paid their respects to the two men, with tributes from as far afield as Australia.	UNB
	But the goal isn’t only to get the reality of homelessness onto social media.	SHAL, AUTH
	“The people of Khyber Pakhtunkhwa are resilient. I did not see hopelessness on any face," he said.	PRES, MERR

Table 5.5: PCL paragraphs correctly classified by the model pre-trained on *Deontology* and missed by the baseline model. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

Some examples are as follows:

- *I am a real estate agent in a large city. So I should make good deals to people who need homes.*
- *I am a high school English teacher. So I should offer tutoring to students who need help.*
- *I am a nun in the Catholic church. So I should pray for people who need it.*
- *I am a camp counselor at a small camp. So I should make sure children are kept safe.*
- *I am a babysitter that works on weekends. So I should ensure the children’s safety.*

The model pre-trained on *deontology*, therefore, learns about what is the right or wrong thing to do in different situations. Examples of PCL often have a similar message, as can be seen in the examples in Table 5.5 for the *deontology* pre-trained model.

The strong results for **Hate** are to some extent surprising, as the style of the tweets in this dataset, which is often about insulting and aggressively addressing people,

is very different from PCL, which is more about praising and pitying individuals or communities. However, the vulnerable communities from the PCL dataset are commonly targeted in hate speech. A model which is pre-trained on hate speech can thus learn about what kind of attitudes towards these communities are acceptable. Moreover, the authoritarian or aggressive tone, the hyperboles and the abuse of adjectives that can be found in hate speech are also common traits of PCL, especially for the categories AUTH, COMP and MET. Some examples of sentences from the Hate dataset are as follows, with the first two being positive examples of hate speech and the last two being negative examples.

- *@user Coward Cameron go on welcome migrants with housing etc while destroying disabled peoples benefits its not a secret ur no good.*
- *Prevent new refugee crisis? You can stop doing the lies n propagandas bullshit.You can't even take care of your poor ppl at home. Space Force is too expensive for the ppl w 2 jobs.You can't even take care of Puerto Rico. Good night millions of homeless on the streets of US.*
- *Why we need to protect refugees from the ideas designed to save them.*
- *Lots of events coming up next week. Sign up to take action! On Aug 15th call Governor Wolf and demand he take action to protect immigrant families. Stop being complicit with Trump/ICE. Governor Tom Wolf...*

Some parallels with the examples for *hate* in Table 5.6 can be observed. First, in the examples above, we see how vulnerable communities are presented as being in need of protection and attention, which is similar to the examples from the PCL dataset in Table 5.6. The authoritarian and aggressive tone from the two positive examples above also resembles the last example for *hate* in Table 5.6.

The **Offensive** dataset presents an interesting case. Although offensive language is inherently harmful, it might be more subtle and indirect than hate speech, which is

Model	Text	Categories
Hate	Apparently in Dr. Ablow's eyes, people who undergo the transgendered process are broken individuals, in need of repair. There are no transgendered people – only people who are confused and in need of treatment to alleviate their condition.	PRES, MET, COMP
	School for the blind, deaf and dumb, Isulo, Anambra State, which parades a number of beautiful structures, is one of the schools battling with lack of facilities to meet the special educational needs of the children. According to Felix Nwaochi, President-General of Isulo Community, the school is seriously in need of water supply as many of the blind students have to fetch water from a stream to survive in the school.	UNB, SHAL, MET
	"I and my daughter Monica are excited about providing a space for disabled people to be able to get together and earn fair prices for their work," Mr. Rogers said.	UNB
	As Maas put it, "the loss of this organisation could unleash an uncontrollable chain reaction. "Kids would be pushed from Unrwa classrooms onto the streets, where they would be more vulnerable to dangerous scenarios such as recruitment efforts by terrorists, who will surely jump at the chance to argue that if we can't keep our aid promises, peaceful coexistence with the West is impossible. Child marriage, child labour, and child trafficking would rise. A generation of children and young people would be lost, in a region more unstable than ever.	UNB, AUTH, MET, COMP

Table 5.6: PCL paragraphs correctly classified by the model pre-trained on *Hate* and missed by the baseline model. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

also a characteristic of PCL. We could also claim that an offensive message might not be entirely ill-intentioned, following the work of Yin and Zubiaga [185], and that the offence might depend more on the receiver of the message than on the author of it. In the Offensive dataset we find then a softer language, where the offensiveness often relies on sarcasm. In other messages, we consider that a pretended positive message towards a person or community unveils an unfair treatment or a generalization that can be considered offensive, which is a definition that would also fit a condescending message, especially for the categories of UNB and MERR. In addition, the communities more susceptible to receiving condescending treatment are often also targeted in offensive discourse. The strongly opinionated texts and the use of the imperative mode, which are often representative of tweets, might also

help the model with categories such as AUTH and PRES. To illustrate these points, we include some sentences extracted from the Offensive dataset, with the first three being negative cases and the last two being positive examples of offensive language.

- *@user @user You care about black men dying at the hands of racist cops? You care about the immigrants being kidnapped at the border? You care about the 3k Puerto Ricans that died? I figured you would bring up antifa. That says everything about you.*
- *@user me and my girl walked through San Francisco and I came up with a new game. Homeless or antifa. The homeless are so much more pleasant.*
- *@user @user @user @user So handsome you are! I think people should donate the shoes to our homeless veterans and other homeless Americans, after they sew a no""""""" circle over the emblem.""""""""*
- *@user @user @user Because some people are bastards. You're hopelessly naïve if you think plenty of frauds won't take advantage of this law change for malign purposes.*
- *@user @user You are name calling a decorated disabled veteran. - shame on you. Know your subject before making slanderous remarks.*

Table 5.7 presents some examples of PCL instances correctly classified by the *offensive* model. We can identify here some of the features of offensive language, such as the treatment of some communities as in need of *our* protection and action. Also, we can see authoritarian attitudes and suggestions of solutions, which share the strongly opinionated discourse of tweets in general and offensive language in particular. Moreover, the last two examples show the type of discourse that, although trying to praise someone, might be offensive because of the assumptions and generalizations which can be inferred from the message.

Model	Text	Categories
Offensive	"We have to sit down, dialogue with those who are agitating and start looking at meaningful solutions that can give them hope. Once a country makes her people to develop a sense of hopelessness, the people will agitate a lot."	UNB, AUTH
	Hundreds of thousands of internally displaced persons (IDPs) belonging to FATA are languishing in refugee camps since the military operations started in the region. Rehabilitation of these people should be the utmost priority of the government. For that purpose, construction of health and education facilities as well as other infrastructure is necessary. According to the committee recommendations, foreign donors for the rehabilitations process could not be approached without legal reforms in FCR.	UNB, COMP
	Who blame for this issue?? The system itself or people? Must be the people's fault for being refugees or being poor, instead born being in the right country.	COMP
	Even people who are disabled can still practice karate and have a sense of accomplishment. It matters not what your state of being is, people are encouraged to excel.	UNB, COMP
	If only we had more stories that championed the brilliance of migrant workers perhaps we 'd be able to challenge the silence that permits them to be treated in such a disdainful way.	UNB, COMP, MERR

Table 5.7: PCL paragraphs correctly classified by the model pre-trained on *Offensive* and missed by the baseline model. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

The relevance of **Irony** may seem less clear than that of the other datasets. However, our experimental results nonetheless show that pre-training on this dataset is beneficial. To understand why this is the case, it is worth pointing out that instances from this dataset often contain strongly opinionated language and value judgments, which are related to the *AUTH* and *PRES* categories, as well as generalizations and hyperboles, which are relevant for the *MERR* and *PRES* categories. Moreover, a speaker using irony often decorates their language with unnecessary, flowery wording, which is relevant for the *MET* category. The following examples from the *irony* dataset illustrate these points:

- *"Now that i can seem to afford good things, material things in life ... its the simple things that i need and really want ... of my life"*

Model	Text	Categories
	As a matter of life views, migrants generally see opportunities where locals don't. they see how their home society has handled different problems and they can draw from that experience to simply copy and paste amazing solutions that change a society. These innovations are what an economy needs to grow and solve its own issues in dynamic ways.	PRES, MERR
Irony	"It 's not just a matter of income poverty. What matters is children in very poor families in crowded, cold and damp houses. There is an income issue, there is a housing supply issue and there is a housing quality issue."	AUTH, COMP
	Bombarded by schizophrenia, addiction and homelessness, you might say that Eoghan O'Driscoll has been to hell and back. but he is finding a new balance through painting. Interview: Michael Lanigan	MET, COMP
	Many celebrities wore blue ribbons to support the American Civil Liberties Union, which is seeking to shed light on the plight of young immigrants facing the potential of being deported.	UNB, SHAL, MET, COMP

Table 5.8: PCL paragraphs correctly classified by the models pre-trained on *Irony* and missed by the baseline model. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

- *@user try having no internet for a month. Now I know how Ethiopians feel.*
- *"so, sane peoples would talk to themselves in twitter because they can't find other sane humans to talk to. that #retweet #ifagree"*
- *"@user I don't think, I know x"*

The model seems to learn from the assumptions, exaggerations and generalizations in the *Irony* dataset, as we can also find them in the *irony* examples in Table 5.8, for instance in the generalization and assumption that migrants see what locals do not. In other examples from Table 5.8, we can see a dichotomy between a dramatic situation and a shallow solution (e.g. painting or wearing blue ribbons), which is reminiscent of the dichotomies that often appear in ironic language. The authoritarian, confident tone of the last two examples extracted from the *Irony* dataset is also a common feature of PCL.

Pre-training on **Sentiment** also improves the model's performance on PCL detection. There are several features in the *Sentiment* dataset which can help the model to detect condescension. For instance, the inputs from this dataset often contain a confident, strongly opinionated tone, characteristic of tweets, which is also a feature of the *AUTH* and *PRES* categories in PCL. To express sentiment, the texts also contain a fair number of adjectives, which can be easily linked to the *COMP* category in PCL. If we look at some examples from the dataset, we can also see a recursive structure of content, where someone does something for another person, a structure also shared by the *UNB* and *SHAL* categories of PCL. Some of these features can be observed in the examples below, extracted from the *Sentiment* dataset:

- *We've got the info on how YOU can help those in need in SLC w/ @user & @user #ad*
- *Support CEO Keith Bradshaw as he spends a night sleeping at Adelaide Oval on THURSDAY raising money for the homeless*
- *'Knock Knock: Live:' David Beckham Surprises Family In Need: Tuesday marked the debut of "Knock Knock:... #family"*
- *"Jeff Foxworthy leads a Bible study with homeless guys on Tuesday mornings, and has for years. How cool is that?"*
- *"In the Oregon experiment, 10,000 previously-excluded people (poor & childless) were given access to Medicaid for the first time"*

The language in the above examples clearly shows unbalanced relations between those who can help and those who are helped, a highly indicative feature of PCL. Furthermore, in these examples, those in a more powerful situation are praised by their charitable actions, which is as well a common theme in PCL, as shown in most of the examples of *Sentiment* in Table 5.9. There, the individuals who help and their actions take center stage in the paragraph, above the community or individuals

Model	Text	Categories
Sentiment	A kind-hearted woman has rescued a 11-year-old girl fleeing from her home in the Sri Lankan refugee camp near Madurai and re-united her with her family with the help of police in Tiruchi.	UNB
	The actor, who will be seen later this month in Avengers: Infinity War, found himself called upon to make the day of a young fan in need. On Wednesday, he hung out with Jacob Monday, who is a 16-year-old from upstate New York who has terminal cancer. The teen, who has a rare form of bone cancer, has a bucket list he's working through and it included meeting his favorite movie star.	UNB, SHAL
	Discrimination of the disabled by society is one of the major problems undermining the progress of democratic practice in the country. It is always the dream of people with disabilities that so long as the disability bill is passed, their position in society will be influenced positively.	PRES, AUTH
	He said the victims who are currently rendered homeless can now be relieved of troubles as the 5,000 iron sheets from Mwanza had arrived, with 1,200 already distributed to victims in Bukoba Municipality.	UNB, SHAL, AUTH
	The boxers were from poor families and had nothing. I was trying to feed them in my own home, and I wasn't thinking about my own family. All I knew was I had food in my house and I had to feed the boxers.	UNB, AUTH, COMP

Table 5.9: PCL paragraphs correctly classified by the model pre-trained on *Sentiment* and missed by the baseline model. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

who receive the action. By pre-training on *sentiment* the model seems to learn associations between some communities and their positions of power and need, and that helping others is considered an action with positive sentiment. This knowledge helps the model to better identify PCL.

5.4 Summary

In this chapter, we explored how different auxiliary tasks can help detect PCL. To this end, we used three strategies for pre-training PCL detection models on ten text classification tasks whose associated data could be related, to some extent, to condescending language.

Regarding the pre-training strategies, we showed how, in this setting, the use of adapters works better than full fine-tuning a pre-trained LM, as the *catastrophic forgetting* that seems to occur while fine-tuning impacts negatively on the PCL detection results.

Focusing on adapters, we find that performance gains are indeed possible in this way, in particular when pre-training on tasks related to harmful language and commonsense morality, which supports the idea that PCL detection requires an assessment of human values. We also found irony detection and sentiment analysis to be useful pre-training tasks. While these tasks are conceptually rather different from PCL detection, we found several similarities in the underlying discourse, such as the use of hyperboles, strongly opinionated language or the abuse of adjectives and flowery wording. In contrast, for tasks focusing on political speech, social justice and stereotypes no improvements were witnessed.

These findings improve our understanding of the nature of PCL, although further research in this area is needed to assess the relation of PCL to other kinds of discourse and the gains which could be obtained by infusing previous related knowledge into a PCL detection model. For instance, stereotypes are a distinctive feature of PCL, despite the disappointing results for the *Stereoset* model. These results, therefore, might be explained by the limited relationship between the communities represented in both datasets and the stereotypes contained in them.

Moreover, each auxiliary task might need specific hyperparameter tuning in order to obtain the best results on PCL detection. However, the objective of this chapter was to obtain a better understanding of the nature and features of PCL by offering an overview of how different types of discourse might be, or not, related to PCL and how we can use these insights for further research.

In the next chapter, we will introduce a new test set for the DPM! dataset and will present the shared task on PCL detection and categorization hosted at SemEval-2022. We will analyze the best approaches and will summarize their most important

insights. We will also share some of the lessons learnt from sharing this challenge with the community.

Chapter 6

Shared Task on Detecting and Categorizing PCL

6.1 Introduction

In this chapter we present an overview of the shared task that we organized on Patronizing and Condescending Language Detection and which was hosted at SemEval-2022. The task attracted the participation of 77 teams during the official duration of the competition, and it is still being used for research purposes and as a learning resource at several universities.

For this shared task, two sub-tasks were considered, in line with the two problems of PCL detection and categorization that we proposed in Chapter 4:

- **Subtask 1** was presented as a binary classification problem, where participants needed to classify a given paragraph as containing PCL or not (i.e., give a label of 0, if the paragraph does not contain PCL, or 1 if it is a positive case of condescension).
- **Subtask 2** was a multi-label classification task, where participants needed to identify which types of PCL are present in each paragraph, if any. In this case, participants had to assign a set of 7 labels to each paragraph, each one of them being 0 or 1 depending on which PCL category is present in the input

text. The categories are based on the taxonomy of PCL categories presented in Chapter 3.

In the next pages, we first provide an overview of the data used for this challenge in Section 6.2. Next, we describe how the task was organized in Section 6.3. After this, in Section 6.4, we discuss the techniques that were employed by the different participants and present the best performing models. To finalize, we summarize some of the insights and lessons learned with this task in Section 6.5.

6.2 Dataset

The seed material for this task is *Don't Patronize Me!* (DPM!), an annotated dataset with Patronizing and Condescending Language towards vulnerable communities, which was introduced in Chapter 3 and 4. This dataset contains 10,467 paragraphs, which were used as the training set for the SemEval task. To create the test set for this task, we annotated 3,832 additional paragraphs, following the same process as in the training set, which was explained in Chapter 3. The next two subsections present statistics of the data contained in the test set, as well as baseline results for the tasks of identifying and classifying PCL, where the models are trained on the entire training set and tested on the test set. From this point, we will refer to the two sets of data as (DPM!) training set and (DPM!) test set.

6.2.1 The DPM! test set at a glance

As in the training set, all paragraphs for the DPM! test set were extracted from news stories from media in twenty English speaking countries, originally provided by the News on Web (NoW) corpus [34]. We also followed the same process of curation. For the annotation, we recruited two new annotators, who annotated the binary

version of the dataset with labels 0-4 (see Section 3.4). The third annotator, who had participated as a referee for the training data, annotated which categories were present in the positive cases.

	dis	hom	hop	imm	mig	need	poor	ref	vul	wom	**Total**
Australia	17	24	22	19	17	18	20	20	15	20	192
Bangladesh	24	17	19	23	19	23	15	23	19	21	203
Canada	18	22	23	22	28	23	18	19	14	23	210
Ghana	13	20	18	18	12	23	4	21	21	20	170
Hong Kong	14	15	16	21	16	20	7	24	21	14	168
India	9	23	13	15	22	18	22	16	15	25	178
Ireland	14	24	20	17	17	17	15	17	27	19	187
Jamaica	22	12	27	19	22	17	6	21	24	23	193
Kenya	23	24	18	19	21	23	19	25	17	14	203
Malaysia	17	27	28	21	17	13	20	17	15	17	192
New Zealand	10	24	14	23	18	24	22	25	23	28	211
Nigeria	20	14	26	22	19	21	22	19	13	20	196
Pakistan	24	18	24	24	18	16	15	19	21	17	196
Philippines	13	19	19	27	16	21	21	23	19	22	200
Singapore	23	19	17	18	15	16	20	28	20	25	201
South Africa	15	21	10	17	19	19	15	25	27	19	187
Sri Lanka	22	18	17	16	18	25	8	19	26	23	192
Tanzania	0	16	6	15	18	23	17	26	18	22	161
UK	20	24	28	20	19	19	15	16	17	23	201
United States	14	13	20	24	20	20	22	13	28	17	191
Total	332	394	385	400	371	399	323	416	400	412	3832

Table 6.1: Number of paragraphs per keyword and country in the test set

In Table 6.1 we show the distribution of paragraphs by country and community in the test set. For some combinations of country-keyword, the natural distribution of the data prevented a totally balanced dataset. For instance, we could not find any paragraph which mentioned the keyword *disabled* in Tanzania's media in our test set, which is coherent with the low number of instances for the same country-keyword combination in the training set, as can be seen in Table 4.1. Similarly, the keyword *poor families* is hardly mentioned in some countries, which makes this keyword one of the least populated in the dataset.

From the 3,832 paragraphs of the test set, 317 were annotated with labels 3 and 4, as containing PCL. For the test set we excluded paragraphs annotated with the label 2, as these are consider to be borderline cases. Those 317 positive paragraphs

	unb	shal	pre	aut	met	com	merr	Total
Dis	19 (48.7%)	3 (7.7%)	0 (0%)	5 (12.8%)	2 (5.1%)	6 (15.4%)	4 (10.3%)	39
Hom	42 (39.6%)	19 (17.9%)	8 (7.5%)	6 (5.7%)	12 (11.3%)	16 (15.1%)	3 (2.8%)	106
Hop	17 (20.5%)	0 (0%)	2 (2.4%)	18 (21.7%)	9 (10.8%)	35 (42.2%)	2 (2.4%)	83
Imm	3 (15%)	0 (0%)	0 (0%)	4 (20%)	1 (5%)	10 (50%)	2 (10%)	20
Mig	7 (43.8%)	0 (0%)	1 (6.3%)	0 (0%)	0 (0%)	6 (37.5%)	2 (12.5%)	16
Need	39 (63.9%)	4 (6.6%)	0 (0%)	5 (8.2%)	2 (3.3%)	10 (16.4%)	1 (1.6%)	61
Fam	36(32.7%)	7 (6.4%)	16 (14.5%)	20 (18.2%)	3 (2.7%)	25 (22.7%)	3 (2.7%)	110
Ref	14 (34.1%)	6 (14.6%)	0 (0%)	6 (14.6%)	2 (4.9%)	11 (26.8%)	2 (4.9%)	41
Vul	14 (53.8%)	1 (3.8%)	0 (0%)	4 (15.4%)	2 (7.7%)	5(19.2%)	0 (0%)	26
Wom	16 (53.3%)	2 (6.7%)	0 (0%)	6 (20%)	0 (0%)	4 (13.3%)	2 (6.7%)	30
Total	207 (38.9%)	42 (7.9%)	27 (5.1%)	74 (13.9%)	33 (6.2%)	128 (24.1%)	21 (3.9%)	532

Table 6.2: Number and % of text spans that have been labelled with each of the PCL categories in the test set, per keyword. The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

contained, in turn, 532 spans annotated with PCL categories. Table 6.2 shows the distribution of categories by keyword in the test set. Similarly to the distribution in the training set, we can see how *Unbalanced power relations* and *Compassion* are the most frequent in the test set. In this data, media expresses unbalanced power relations especially towards people in need, but also towards vulnerable groups and women. We often see expressions of compassion towards immigrants, those considered hopeless and migrants, which follows the trends that we saw at the analysis of the training set. *The poorer, the merrier* also repeats as the least recurrent category. While we see a similar distribution between the training and the test set in general, some combinations of keyword-category do not present any instance in the test set, due to its smaller size. For instance, comparatively, presuppositions are scarce in the test set, with eight out of ten communities presenting none or very few examples of this category.

Regarding the analyzed communities, *poor families* and *homeless* receive more condescending treatment, whereas *migrants*, *immigrants* and *women* present less examples of PCL, which shows the same trends as the training set.

6.2.2 DPM! test set baseline results

In order to keep testing the DPM! dataset as a linguistic resource for further research in modeling PCL, we replicate the experiments from Chapter 4, using the entire training set for training (i.e. the full 10467 paragraphs) and the new test set for testing.

We experiment with the same models, namely Support Vector Machine models based both on Bag of Words and on word vectors, Bidirectional LSTM, a random classifier and different Language Models from the BERT family. We run experiments in the settings of Subtask 1 and Subtask 2, namely binary and multi-label classification. The specifications for each one of the models are explained in Section 4.3.1.

In Table 6.3, we see the baseline results for Subtask 1, where the model has to classify a paragraph as containing or not PCL in a binary classification problem. As in the previous experiments, we show the average of 5 runs for each one of the models. The results replicate most of the trends seen in Table 4.4, corresponding to the experiments in the training set, although on the test set, RoBERTa-base obtains the best results, followed by BERT-large. The performance of RoBERTa-large is again highly inconsistent. The SVM based on word embeddings still obtains fairly good results, surpassing the BiLSTM approach and RoBERTa-large, due to the aforementioned inconsistency of this last model.

	P	R	F1
Random	8.79	53.00	15.08
SVM-BoW	32.16	34.70	33.38
SVM-WV	36.24	54.55	43.55
BiLSTM	34.94	50.73	41.01
DistilBERT	36.45	65.49	46.83
BERT-base	40.67	70.03	51.44
BERT-large	41.62	73.63	53.15
RoBERTa-base	**42.65**	**73.82**	**54.05**
RoBERTa-large	33.47	57.16	42.19

Table 6.3: Results on the test set for the problem of detecting PCL, viewed as a binary classification problem (Subtask 1).

Table 6.4 shows the performance of the same models in Subtask 2, seen as a multi-label classification task. In these experiments, we introduce a new setting, i.e., we also include negative examples of PCL, in order to make the task more challenging and also more realistic. This is the same setting we proposed for Subtask 2 at the SemEval-2022 shared task. The results show that different models perform better for different categories, with RoBERTa-large ouperforming the other models in *UNB*, *MET* and *COMP*; BERT-base being the best one at detecting the presence of *SHAL*, *PRES* and *AUTH*, and RoBERTa-base being specially good at the detection of *MERR*. As in the experiments in the training set, the models based on SVMs obtain fairly good results, specially for *MERR*, which stands as the most difficult category, with most models not being able to predict any instance of it. The results shown in the table are the average of 5 runs.

6.3 Shared Task Setting

In this section, we explain the setting of the SemEval-2022 task on Patronizing and Condescending Language Detection.

Data

Participants were provided with sentences in context (paragraphs), extracted from news articles, in which one or several predefined vulnerable communities are mentioned. The data provided was divided as follows:

- **Training data**: The 10,467 annotated paragraphs from the DPM! training set were provided as training data. To frame Subtask 1 as a binary classification problem, paragraphs with labels 0 and 1 were considered as negative

	Random			**SVM-BoW**			**SVM-WV**		
	P	R	F1	P	R	F1	P	R	F1
Unb	5.55	47.51	9.94	16.58	44.34	24.14	30.06	47.06	36.68
Shal	0.95	41.86	1.86	6.34	30.23	10.48	18.84	30.23	23.21
Pres	1.53	60.42	2.98	3.88	18.75	6.43	13.54	27.08	18.06
Auth	1.82	45.33	3.50	7.63	26.67	11.87	14.89	18.67	16.57
Met	1.02	51.35	2.00	2.39	13.51	4.07	11.76	21.62	15.24
Comp	3.31	49.61	6.21	8.84	32.56	13.91	19.14	37.98	25.45
Merr	0.36	33.33	0.72	16.67	14.29	15.38	16.67	9.52	12.12
	BiLSTM			**DistilBERT**			**BERT-base**		
	P	R	F1	P	R	F1	P	R	F1
Unb	36.13	41.45	38.45	41.66	56.74	48.03	43.21	64.25	51.67
Shal	31.33	6.51	10.36	38.49	39.53	38.94	54.00	49.30	**51.44**
Pres	12.19	5.00	6.52	24.26	26.25	25.17	31.13	34.58	**32.72**
Auth	5.71	0.53	0.98	26.82	17.60	21.19	37.27	27.73	**31.77**
Met	1.25	1.08	1.16	26.32	18.38	21.53	22.18	14.59	17.57
Comp	26.06	27.29	26.26	32.04	46.67	37.97	34.66	53.95	42.19
Merr	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	BERT-large			**RoBERTa-base**			**RoBERTa-large**		
	P	R	F1	P	R	F1	P	R	F1
Unb	42.45	71.58	53.29	44.99	71.22	55.09	44.94	72.58	**55.46**
Shal.	37.90	47.91	42.19	36.77	55.81	44.32	35.40	60.93	44.72
Pres	24.48	47.50	32.16	22.24	31.67	26.08	22.93	42.92	29.87
Auth	30.07	24.53	26.74	31.76	25.87	28.48	28.65	32.00	30.02
Met	28.35	20.00	23.30	35.67	31.35	33.36	36.15	38.38	**36.97**
Comp	32.06	61.09	42.02	36.16	59.22	44.88	36.28	61.86	**45.71**
Merr	6.67	0.95	1.67	54.45	19.05	**27.89**	37.89	18.10	24.45

Table 6.4: Results on the test set for the problem of categorizing PCL, viewed as a paragraph-level multi-label classification problem (Subtask 2). The categories are abbreviated as follows: unbalanced power relations (unb), shallow solution (shal), presupposition (pre), authority voice (auth), metaphors (met), compassion (comp), the poorer, the merrier (merr).

examples, while paragraphs with labels 2, 3 and 4 were considered as positive examples of PCL. The original labels on the scale from 0 to 4 were also made available. The 993 positive examples in the training data are labelled with the corresponding PCL categories. The meta-information for each paragraph includes an article id, a paragraph id, the country where the article was published, the community mentioned in the paragraph and the binary or multi-label annotation. For the category-level annotated data, span annotations for each category were also provided.

- **Test data**: The 3,832 paragraphs from the DPM! test set were released as

test set for the task, with the same format and meta-information as the training set. For the sake of the shared competition, however, nor labels or span annotations were provided for the test set.

We welcomed the use of external resources in this task. Participants were encouraged to explore transfer learning or data augmentation techniques with a variety of source corpora and language resources.

Evaluation

To rank the system submissions we used the following evaluation metrics:

- **Subtask 1**: F1 score for the positive class.
- **Subtask 2**: Macro-averaged F1 over all categories.

For both subtasks, we compared all the system submissions results with the rest of participating systems and with the task official baseline. Specifically, we ran RoBERTa-base for binary and multi-label classification for Subtask 1 and Subtask 2, respectively, trained with default parameters for 1 epoch on the full training set.

Participation Framework

The task was hosted on CodaLab[1], with participants needing to register and submitting their results through the platform. The competition involved the following three phases:

- **Practice phase:** The 10,467 paragraphs from the training data were split into 8,373 training paragraphs and 2,094 validation paragraphs. This was done

[1]https://competitions.codalab.org/competitions/34344

to allow participants to compare their systems on a public leader board. The training-validation split respected the natural distribution of labels in the data.

- **Evaluation phase:** This was the official evaluation phase for the SemEval competition. The test data was released and the leader board for this phase remained hidden to prevent participants from fine-tuning their systems on the test data. Each participant was allowed two different submissions for each subtask.

- **Post-evaluation phase:** The learderboard for the evaluation phase and the official ranking for each subtask were published, as the SemEval competition ended. Participation in the SemEval task is no longer possible, but the competition remains open on CodaLab to allow participants to re-test and further improve their systems.

6.4 Results and Discussion

A total of 77 different teams participated in the evaluation phase of our task, with 145 valid submissions for Subtask 1 and 84 for Subtask 2. For the competition, we allowed a maximum of 2 submissions per team. A total of 42 out of 77 teams outperformed the baseline for Subtask 1, while 37 out of 48 outperformed the baseline for Subtask 2. Tables 6.5 and 6.6 present the rankings for Subtasks 1 and 2, respectively, where we have only listed the best-performing system for each team. For Subtask 1 (Table 6.5), we report Precision (%), Recall (%) and F1 score (%) for the positive class. For Subtask 2 (Table 6.6), we report F1 score (%) for each one of the categories and the macro-averaged F1 score (%) for all categories.

For Subtask 1, the best-performing systems used the following strategies:

Team PALI-NLP [78] used an ensemble of pre-trained RoBERTa models [98]. While training, they applied grouped Layer-Wise Learning Rate Decay [77], based

on the idea that different layers capture different types of information [186]. By optimizing the learning rate in different layers, the model captures more diverse and fine-grained linguistic features of PCL. To tackle the class imbalance in the dataset, they use weighted random samples [70] to emphasize the positive instances.

Team STCE created adversarial examples to train an ensemble model of RoBERTa and DeBERTa [71]. They also used weighted samples to address the class imbalance and explored different loss functions, establishing Cross Entropy and the contrastive loss algorithm NT-Xent introduced by Chen et al. [26] as first and second loss function, respectively.

For Subtask 2, the best-performing systems used the following strategies:

Team BEIKE NLP [35] participated with a system based on prompt learning [128, 18]. They first reformulated PCL detection as a cloze prompt task and then fine-tuned a pre-trained DeBERTa model.

Team PINGAN Omini-Sinitic [174] proposed an ensemble model which used prompt training and a label attention mechanism, by adding a new label-wise attention layer ([41, 172]. Their system over-samples the positive examples. They also use a form of transfer learning from Subtask 1 to Subtask 2, by pre-training on Subtask 1 and using the resulting model as the starting point for training a model for Subtask 2.

For both sub-tasks, unsurprisingly, most systems relied on pre-trained language models, although a few teams used CNN, LSTM, SVM or Logistic Regression based systems (Xu [184], PC1, I2C [138], Ryan Wang, McRock [156], Amrita_CEN [56], SATLab [11] and Team Lego [158], among others), or an ensemble of some of the above together with Language Models (UTSA NLP [194], Taygete [27]). Although the use of Language Models usually outperformed other systems in this task, some

	TEAM	P	R	F1		TEAM	P	R	F1		TEAM	P	R	F1
1	PALI-NLP	64.6	65.6	65.1	**27**	ML_LTU	58.0	51.4	54.5	**53**	RNRE NLP	39.0	50.2	43.2
2	stce	63.3	66.9	65.0	**28**	ZYBank-AI	54.8	53.9	54.4	**54**	SATLab	34.8	55.2	42.7
3	ymf924	63.8	65.6	64.7	**29**	Team LRL_NC	60.7	49.2	54.4	**55**	J.U.S.T-DL	49.0	37.5	42.5
4	BEIKE NLP	61.2	67.2	64.1	**30**	CS-UM6P & ESL	55.2	53.3	54.3	**56**	MaChAmp	58.8	32.8	42.1
5	holdon	60.3	67.5	63.7	**31**	Felix&Julia	40.1	77.3	52.8	**57**	I2C	61.1	31.2	41.3
6	cnxup	62.7	64.7	63.7	**32**	Stanford ACM	40.2	76.7	52.7	**58**	SMAZ	36.3	47.6	41.2
7	abcxyzw	58.8	68.5	63.3	**33**	UtrechtUni	44.6	62.5	52.0	**59**	MASZ	36.3	47.6	41.2
8	nowcoder	58.2	68.5	62.9	**34**	CSECU-DSG	59.0	46.4	51.9	**60**	Amrita_CEN	32.2	52.1	39.8
9	PINGAN Omini-Sinitic	61.8	63.7	62.7	**35**	Sapphire	59.4	46.1	51.9	**61**	Anonymus	27.6	59.9	37.8
10	bigemo	57.1	69.4	62.7	**36**	Ablimet	61.5	44.8	51.8	**62**	matan-bert	35.4	40.4	37.7
11	Leo_team	60.1	64.0	62.0	**37**	SSN_NLP_MLRG	42.3	66.6	51.7	**63**	Team LEGO	24.8	56.5	34.5
12	PAI-Team	66.3	57.7	61.7	**38**	Team PiCkLe	46.0	58.0	51.3	**64**	TüSoXi	38.8	29.3	33.4
13	Anonymus	53.5	70.4	60.8	**39**	sua	54.0	48.6	51.2	**65**	RNRE NLP RFC	30.0	36.9	33.1
14	BLING	63.5	55.5	59.3	**40**	UCL xNSI	41.5	65.3	50.7	**66**	jct_meir	25.3	47.0	32.9
15	Taygete	53.6	66.3	59.2	**41**	MS@IW	50.2	51.1	50.6	**67**	isys	22.4	59.3	32.5
16	NLP-Commonsense Reasoning	61.2	56.8	58.9	**42**	Univ. of Bucharest Team	49.1	50.8	49.9	**68**	AliEdalat team	18.4	87.1	30.3
17	GUTS	61.3	54.9	57.9	**43**	**RoBERTa Baseline**	**39.4**	**65.3**	**49.1**	**69**	McRock	23.4	39.1	29.3
18	DH-FBK	64.2	52.7	57.9	**44**	rematchka	44.5	53.9	48.8	**70**	Waad	64.0	18.0	28.1
19	ULFRI	56.4	58.7	57.5	**45**	fengxing	63.8	39.4	48.7	**71**	Ryan Wang	17.0	60.9	26.6
20	TUG-CIC	60.2	54.9	57.4	**46**	flerynn	67.2	38.2	48.7	**72**	PC1	37.8	18.6	25.0
21	amsqr	54.8	59.9	57.2	**47**	Team YNU-HPCC	65.9	36.6	47.1	**73**	UTSA_NLP	14.0	35.0	20.0
22	UMass PCL	52.9	58.4	55.5	**48**	niksss	51.8	42.0	46.3	**74**	yaakov	11.2	10.1	10.6
23	LastResort	51.5	59.9	55.4	**49**	JustTeam	55.0	39.8	46.2	**75**	ilan	14.5	6.0	8.5
24	Team Double_A	47.2	66.6	55.2	**50**	BWQ	51.0	41.3	45.6	**76**	Jiaaaaaa	8.2	6.3	7.1
25	thetundramanagainstpcl	54.3	55.5	54.9	**51**	Tesla	36.0	57.7	44.3	**77**	Anonymus	29.7	3.5	6.2
26	Xu	46.2	66.9	54.6	**52**	ASRtrans	35.6	58.4	44.2	**78**	Anonymus	10.6	2.8	4.5

Table 6.5: Ranking by teams for Subtask 1 at SemEval-2022 shared task: Binary Classification

	TEAM	UNB	SHAL	PRES	AUTH	MET	COMP	MERR	Avg
1	BEIKE NLP	65.6	52.9	36.9	40.7	35.9	49.2	47.1	46.9
2	PINGAN Omini-Sinitic	59.7	53.1	41.7	43.4	42.7	51.3	15.4	43.9
3	PAI_Team	57.6	45.2	35.2	39.4	38.4	44.5	26.7	41.0
4	stce	62.2	54.8	38.1	32.8	33.3	51.0	8.7	40.1
5	PALI-NLP	61.8	54.1	37.7	32.8	32.8	51.2	8.7	39.9
6	Leo_team	57.3	47.0	28.8	36.1	34.8	47.4	27.0	39.8
7	Anonymus	59.9	49.1	38.5	37.1	35.0	48.6	8.3	39.5
8	ymf924	61.6	54.1	36.8	31.3	33.3	50.0	8.7	39.4
9	bigemo	62.5	56.1	38.0	24.3	31.3	49.4	8.7	38.6
10	holdon	62.2	56.1	32.9	23.1	33.3	48.7	8.7	37.9
11	cnxup	60.2	53.3	30.6	24.1	40.0	48.1	8.7	37.8
12	Taygete	59.7	45.8	33.3	21.8	30.4	53.6	18.8	37.6
13	DH-FBK	52.5	36.2	27.0	37.7	31.9	46.0	30.3	37.4
14	abcxyzw	60.7	53.3	34.5	21.8	32.8	50.0	8.3	37.4
15	nowcoder	59.8	50.0	32.2	22.8	39.4	47.8	8.3	37.2
16	GUTS	55.6	47.4	24.0	34.3	25.6	44.4	27.6	37.0
17	BLING	55.1	38.9	23.4	29.0	31.5	50.9	26.7	36.5
18	UMass PCL	53.9	42.4	29.1	30.7	33.3	40.8	23.5	36.3
19	CS-UM6P & ESL	57.0	42.0	25.7	25.2	20.5	46.8	21.4	34.1
20	Fengxing	46.4	46.3	23.0	26.5	33.3	38.7	24.0	34.0
21	Team LRL_NC	52.1	42.7	25.2	30.4	28.8	43.3	14.8	33.9
22	thetundramanagainstpcl	50.5	50.0	18.4	16.5	20.3	41.5	24.0	31.6
23	Xu	55.0	48.4	28.0	24.0	13.6	49.0	0.0	31.1
24	SATLab	42.4	33.1	17.0	23.2	17.5	31.5	14.2	25.6
25	Felix&Julia	36.6	35.1	17.6	22.1	21.1	28.5	16.7	25.4
26	AliEdalat team	53.9	37.7	25.6	26.2	13.5	11.3	9.1	25.3
27	Tesla	43.7	38.3	16.3	19.2	17.9	35.7	0.0	24.5
28	Waad	36.9	33.3	17.5	15.3	16.5	28.7	19.5	24.0
29	McRock	32.3	32.9	19.2	20.6	22.2	26.4	7.1	23.0
30	rematchka	37.7	21.4	18.8	21.2	15.5	26.1	13.0	22.0
31	Team Double_A	33.5	31.9	18.4	19.1	23.4	24.5	0.0	21.5
32	SSN_NLP_MLRG	34.6	33.8	20.7	19.3	12.1	27.7	0.0	21.2
33	ASRtrans	18.6	8.8	8.3	19.8	13.2	27.8	35.7	18.9
34	MaChAmp	30.4	21.3	3.6	10.9	30.8	5.0	6.3	15.5
35	Team PiCkLe	10.9	22.5	14.4	21.0	19.2	6.5	11.5	15.2
36	LastResort	15.8	24.8	10.0	9.3	16.0	11.3	14.8	14.6
37	Ablimet	12.6	14.1	6.5	7.2	14.0	17.2	17.1	12.7
38	**RoBERTa Baseline**	**35.4**	**0.0**	**16.7**	**0.0**	**0.0**	**20.9**	**0.0**	**10.4**
39	BWQ	16.0	12.5	7.2	9.7	7.0	11.4	3.9	9.7
40	Stanford ACM	16.0	26.5	4.2	0.0	0.0	8.6	12.1	9.6
41	Team LEGO	11.8	20.6	1.9	6.4	6.5	10.2	0.0	8.2
42	CSECU-DSG	33.4	0.0	0.0	0.0	0.0	21.8	0.0	7.9
43	Univ. of Bucharest Team	14.8	21.7	3.5	0.0	3.9	8.3	0.0	7.4
44	PC1	11.8	12.0	6.1	8.7	2.6	8.9	0.0	7.2
45	Team YNU-HPCC	10.9	0.8	3.5	3.3	0.0	5.8	0.0	3.5
46	NLP-Commonsense Reasoning	9.7	0.2	0.0	3.2	3.2	4.4	1.1	3.1
47	Jiaaaaaa	2.8	1.9	0.0	2.0	0.0	4.8	6.9	2.6
48	Anonymus	5.9	8.3	0.0	2.4	0.0	1.4	0.0	2.6
49	niksss	0.0	1.0	0.0	0.0	0.0	0.0	1.1	0.3

Table 6.6: Ranking by teams for Subtask 2 at SemEval-2022 shared task: Categories Classification.

LSTM models, such as the one submitted by team Xu [184], achieved competitive results.

The ensembling of different models was also a popular technique. Other strategies that proved successful include adversarial training, data augmentation and multitask learning. In the following, we summarize how these techniques were used by the different systems.

Ensemble learning Ensembling different models has previously been found useful for text classification [117, 87, 47]. Accordingly, ensembling was one of the most common strategies for improving on baseline PCL detection methods. Most of the teams combined different language models (e.g. PALI-NLP [78], STCE, PINGAN Omini-Sinitic [174], PAI_Team, LRL_NC [164], SSN_NLP_MLRG [1], AS-Rtrans [140], amsqr [111], UMass PCL [91]). Considering the choice of language models, the most successful systems either used RoBERTa, DeBERTa or an ensemble which included the former ones and other models. For instance, these models were used by the best performing teams for both subtasks, i.e. PALI-NLP [78] and STCE for Subtask 1 and BEIKE NLP [35] and PINGAN Omini-Sintic [174] for Subtask 2. To fine-tune the language models effectively, incorporating a contrastive loss function, in addition to the standard cross-entropy loss, has also proved useful. Finally, it should be noted that the combination of language models with different types of neural networks (Taygete [27], UTSA NLP [194]) also obtained interesting results.

Balancing class distribution The class imbalance in the dataset was addressed by participating teams in different ways. Some teams opted for downsampling the number of negative examples (Ryan Wang, LastResort [2], MS@IW [106]), while others tried a cost-sensitive learning approach to address this issue (Amrita_CEN [56]). However, the most popular approach to balance the class distribution was through data augmentation (amsqr [111], Xu [184], Utrech Uni, UMass PCL [91], among others). To create new positive examples, participants used strategies such as the use of large generative models like GPT3 [18] or T5 [137] (MS@IW [106], PINGAN Omini-Sinitic [174] and Tesla [12]); back-translation (Taygete [27]); the addition of synonymous sentences to the original data (I2C [138]), or the application of the so-called Easy Data Augmentation methods, a set of simple but effective techniques such as synonym replacement, random insertion, random swap, and random deletion (AliEdalat [43]) [178, 141].

External resources Various types of external resources were used. For example, lexical databases such as WordNet [108] were used to augment, enrich and improve the training data (Ali Edalat [43]). Datasets from related tasks, including TalkDown [175], and two metaphor detection datasets, namely MOH [109] and VUA [160], were used both for pre-training and / or for data augmentation by different teams. PAWS, a dataset with Paraphrase Adversaries from Word Scambling, [189] and xTREME, a bechmark for Cross-Lingual Transfer Evaluation of Multilingual Encoders [79], was also used to improve several systems (Ali Edalat [43], ASRtrans [140], Tesla [12], MaChAmp [167]). Other related NLP challenges served as auxiliary tasks for pre-training PCL models (AliEdalat [43], UMass PCL [91]), although such strategies were not always successful (ULFRI [89]). The MaChAmp [167] team used 7 SemEval-2022 tasks, including ours, for training a model based on multi-task learning. The DH-FBK team [139] also opted for multi-task learning, but they only used the data from the Don't Patronize Me dataset itself to create auxiliary tasks. For instance, they trained their model to predict the uncertainty of a label in Subtask 1, using the fine-grained set of labels (0-4); the agreement of the annotators in Subtask 2; the spans where the categories were present; or the country of origin of the news outlets. AliEdalat [43] similarly used the meta-information from the Don't Patronize Me dataset as additional features for training their model.

Prompt learning Using prompts has also proven useful for PCL detection (BEIKE NLP [35], PINGAN Omini-Sintic [174], Ablimet). Specifically, the teams used prompts such as "*[paragraph]* is *[label]*", or "is *[paragraph] [label]*?" where *[paragraph]* is the original input. For Subtask 1, *[label]* is a natural language description of the binary class label (e.g. *"is (not) condescending or patronizing"*). For Subtask 2, *[label]* is the label of a given PCL category.

6.5 Summary

In this chapter, we have described the shared task on Patronizing and Condescending Language Detection hosted at SemEval-2022. We have presented a new test set for the DPM! dataset and offered an overview of the organization, participation framework, results and insights obtained by sharing this challenge with the community.

Altough PCL detection is a relatively new challenge for the NLP community, the high level of participation in this task has provided the community with valuable new insights about how to tackle this problem. A total of 42 out of 77 teams in Subtask 1 and 37 out of 48 for Subtask 2 outperformed the RoBERTa baseline, which proves that PCL detection and categorization is possible and that improvements on this task can be made from different angles. On the one hand, the performance of the best-performing systems shows that a judicious usage of state-of-the-art text classification techniques can bring significant benefits to PCL detection, especially when it comes to addressing the relative scarcity of the available training data and closely related external resources. On the other hand, other systems also obtained interesting results by exploring alternative approaches, like using related data to improve their models performance or further exploiting the subtelty and subjectivity of PCL, by using the disagreement between annotators (which can be seen as a limitation of the data) as a valuable signal for a better modelling of PCL.

In summary, the gains obtained by the different approaches tell us that there still remains considerable scope for further improvements and that there are different avenues to explore towards a better understanding of PCL. It is our expectation that further improvements may need to rely on techniques that are specifically targeted at PCL, e.g. by exploiting insights from linguistics about the linguistic features of PCL, or by building explicit models of stereotypes of vulnerable communities.

In the next chapter we will try to explain why generic models worked so well in

what has been considered by human annotators as a challenging, very subjective task. We will unveil that there are, in fact, two kinds of PCL, one based on linguistic features and one based on thematic and stereotypical discourse. We will explore to what extent learning about community-related stereotypes might impact models in the classification of condescending messages.

Chapter 7

Identifying Condescending Language: A Tale of Two Distinct Phenomena?

7.1 Introduction

Patronizing and Condescending Language is characterized, among others, by its subtle nature. It thus seems reasonable to assume that detecting condescending language in text would be harder than detecting more explicitly harmful language, such as hate speech. Moreover, identifying PCL often seems to require a deep commonsense understanding of human values, as we explained in Chapter 5. For instance, consider the following example from the DPM! dataset:

> "People across Australia ordered pizzas to be delivered on Saturday night, with the ample leftovers donated to local homeless shelters."

We can understand that, although donating food can be socially valuable, the impact of this particular action is painted in an excessively positive light (e.g. as evident in the phrase *ample leftovers*). In addition, this seems to refer to a campaign to increase the consumption of pizzas with the excuse of helping homeless people,

which as humans we might also find condescending. However, an NLP model might struggle to infer such connotations.

Nevertheless, the results of the SemEval-2022 task devoted to this topic and presented in Chapter 6, paint a different picture. The top-ranked submissions achieved a remarkably strong performance, which seems to somewhat undermine the assumption that the subtle nature of PCL would make its detection inherently hard. Moreover, even the best systems [35, 174, 78] relied on a judicious use of more or less generic text classification techniques, improving on the RoBERTa [98] baseline by addressing the class imbalance, adding a contrastive learning loss, using ensembles of language models, etc. In particular, there was little evidence of the presumed need to focus on commonsense understanding of human values.

In this chapter, we analyze the surprising effectiveness of standard text classification methods in more detail. In particular, we highlight the presence of two rather different types of condescending language in the DPM! dataset. Specifically, we identify that some inputs are condescending because of the way they talk about a particular subject, i.e. condescending language in this case is a linguistic phenomenon, which can, in principle, be learned from training examples. However, other inputs are condescending because of the nature of what is said, rather than the way in which it is expressed, e.g. by emphasizing stereotypes about a given community. In such cases, our ability to detect condescending language, with current methods, largely depends on the presence of similar examples in the training data.

In order to test the aforementioned hypotheses about linguistic and thematic PCL, this chapter presents an analysis of the DPM! dataset. First, we carry out two experiments in which models are trained such that they are prevented, to some extent, from learning about condescending themes associated with individual communities. We then complement these results with a qualitative analysis based on ideas from critical Discourse Analysis (CDA), a technique which emerged from Critical Linguistics in the 1970s [54, 46, 45, 182, 169, 82]. CDA looks at the relation between power

and language, and how discourse expresses social hierarchy and inequalities, as we saw in Chapter 2. This qualitative analysis provides further support for the idea that (i) PCL detection models can identify linguistic PCL even if they have not seen similar cases during training while (ii) their ability to detect instances of thematic PCL is much more dependent on the training examples.

The rest of this chapter is organized as follows: First, in Section 7.2, we define and illustrate what we understand as linguistic and thematic PCL in the context of this thesis. In Section 7.3, we introduce the methodology we follow in this chapter, by explaining our experimental setup in Section 7.3.1 and the process to select community-related terms in Section 7.3.2. After this, we discuss the performance of LMs on PCL detection after omitting community-specific data or community-related terms in Sections 7.4 and 7.5, respectively. After a qualitative analysis of these results in Section 7.6, we summarize the chapter and present the conclusions derived from this work in Section 7.7

7.2 Linguistic PCL and Thematic PCL

Our central argument to explain the good results of the SemEval task is that the DPM! dataset contains examples of two rather distinct types of condescending language, and that the difference between the two is fundamental to understanding why the task, as it has been formulated, might be significantly easier than the task of detecting condescending language in general. We argue that a deeper understanding of these two phenomena might lead to better performance on PCL detection. We will refer to these two types as *linguistic PCL* and *thematic PCL*.

Linguistic PCL. Some instances of PCL are related to the way in which a given claim is expressed. Consider the following example:

> "...we must rally together as humans, understanding that we have a responsibility to help the world's most vulnerable to survive and rebuild their lives [...]"

In this sentence, we can see two common aspects of PCL. First, expressions such as *we must* or *we have a responsibility*, indicate an authority voice and attitude [157]. Second, the sentence evokes the idea of a *saviour* and a *victim*. Note how the condescending tone of the sentence is related to linguistic aspects that are relatively easy to identify (e.g. the presence of modal verbs such as *must*) and largely independent of the community being referred to. We will refer to such cases as *linguistic PCL*. Our hypothesis is that detecting linguistic PCL is relatively straightforward for language models, as this is ultimately about learning to detect a particular writing style [84].

Thematic PCL. There are also examples of PCL where the message itself is condescending, irrespective of how it is formulated. We will refer to such cases as *thematic PCL*. Consider the following example:

> "The problem of what to do about the Dreamers, as the immigrants are known [...]"

Calling young immigrants *Dreamers* has condescending connotations, as it implies that the author is in a privileged position which the immigrants aspire to reach. To recognize this, we need a deeper understanding about the nature of condescending language, and we need access to particular world knowledge. For instance, we need to know that the author refers to the DREAM Act[1] and that this tries to protect young immigrants brought to the US as children and fulfill their aspiration to live in America as a *dreamed life*. Our hypothesis is that detecting thematic PCL often requires a level of understanding about human values, and the world in general, that goes

[1] www.americanimmigrationcouncil.org/research/dream-act-overview

above what we can expect to be captured by standard language models. However, the training and test data used in the SemEval task, namely the DPM! dataset, is focused on a small number of vulnerable communities, with the same communities being covered in the training and test data. As such, the model may detect instances of PCL by identifying that they express a similar argument as some training example, rather than by developing an understanding of the underlying reasons why a given example is condescending. In this case, we can expect the model to fail to detect PCL towards communities that are not seen in the training set. Similarly, the model may struggle to adapt when there is a change in the themes that appear in PCL towards previously seen communities.

7.3 Methodology

This section introduces the methodology we follow in the work presented in this chapter. In Section 7.3.1, we describe the basic experimental setup for our experiments. Next, we introduce a simple strategy for characterizing topics or themes that are strongly associated with particular vulnerable communities or groups in Section 7.3.2.

7.3.1 Experimental Setup

Dataset. For these experiments, we use the entire DPM! dataset, the same that was provided for the Patronizing and Condescending Language Detection task at SemEval-2022. This dataset consists of 14,299 annotated paragraphs (10,467 for training and 3,832 for testing). The DPM! training set is introduced in Chapters 3 and 4, and the test set is introduced in Chapter 6. In the experiments presented in this chapter, we only use the binary labels from the dataset, i.e. whether a paragraph is considered to contain PCL or not. In order to put the data in context,

	Neg. Inst.	Pos. Inst.	% Pos.Inst.
Migrant	1052	36	3.3
Immigrant	1031	30	2.8
Refugee	981	86	8.1
In need	906	176	16.3
Poor fam.	759	150	16.5
Vulnerable	1000	80	7.4
Women	1018	52	4.9
Disabled	947	81	7.9
Homeless	899	178	16.5
Hopeless	881	124	12.3
All data	9474	993	9.5

Table 7.1: Number of negative and positive training examples per community. We also report the percentage of positive instances.

we show the number of positive and negative instances for each community in the training data in Table 7.1 and in the test data in Table 7.2. In both cases, the numbers show a highly unbalanced distribution of positive and negative instances, both in community-specific data and in the entire dataset. However, some communities, such as *homeless*, *poor families* and people *in need* present a much higher percentage of condescending messages than others, like *immigrants*, *migrants* or *women*.

	Neg. Inst.	Pos. Inst.	% Pos. Inst.
Migrant	359	12	3.2
Immigrant	383	17	4.3
Refugee	390	26	6.3
In need	357	42	10.5
Poor fam.	267	56	17.3
Vulnerable	382	18	4.5
Women	390	22	5.3
Disabled	308	24	7.2
Homeless	337	57	14.5
Hopeless	342	43	11.2
All data	3515	317	8.3

Table 7.2: Number of negative and positive test examples per community. We also report the percentage of positive instances.

Training Details For our experiments, we fine-tune RoBERTa-base [98] on different versions of the training set. While better results have been reported for

RoBERTa-large and DeBERTa [78, 35, 174], we found the results with RoBERTa-base to be more stable across different runs, which is more important than the absolute level of performance for the analysis in this work. We train our models for 5 epochs, using the Transformers library [183], where we use the AdamW optimizer with a learning rate of $1e^{-5}$ and a batch size of 4. All the reported results have been averaged over 5 runs. To tackle the class imbalance, we down-sample the negative cases to 5,000 and over-sample the positive cases five times during training.

7.3.2 Extracting Community-Related Terms

We associate each of the vulnerable communities from the DPM! dataset with a set of terms, which essentially describe the topics or themes that are specific to, or at least strongly related to, that community. To associate terms with a given community, we compare the set of paragraphs from the training set which mention that keyword (e.g. *homeless*) with the remaining paragraphs. We first select those terms that are mentioned in at least five paragraphs for the considered community. Then we rank these terms according to Pointwise Mutual Information (PMI), i.e. by comparing how strongly the presence of a given term x (e.g. *addicts*) is associated with the presence of the community keyword y (e.g. *homeless*), as follows:

$$PMI(x;y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{7.1}$$

where $p(x,y)$ is the percentage of paragraphs that contain both x and y, while $p(x)$ is the percentage of paragraphs that contain x, and similar for $p(y)$. Finally, we select the top-k highest ranked terms for each community, where we have considered $k = 25$, $k = 100$ and $k = 500$ in our experiments. Note that the selected terms are not necessarily indicative of PCL. However, even for $k = 25$ we observed that some of the selected terms reflect condescending attitudes.

We identify three types of themes:

a) Themes which are not significant for the purpose of this work, as they are not related to the potentially vulnerable communities we are studying (e.g. *Yankees*, as the team publishes the list of the *disabled* players for the next match, or *SUV*, as a type of vehicle one might be *in need* of). Because of them not being meaningful in this context, we consider that including or removing them will not affect the performance of the model for identifying PCL.

b) Themes which are closely related to the community and can help a LM to recognise which group is being addressed in a text and, eventually, to identify (*thematic*) PCL towards that community.

c) Themes which, although they appear in our selection as associated with a specific community, are actually meaningful for identifying *linguistic PCL* for this or other communities. Removing these themes from the training data might affect negatively the performance of the model. For example, *dignity*, *charity* or *hopeless* are often indicatives of PCL, independently of the community they are referring to.

For illustrative purposes, Table 7.3 shows a selection of terms that were found for $k = 100$.

Finally, we analyse to what extent the ten keywords from the DPM! dataset refer to distinct communities. To this end, we represent each keyword/community as a PMI-weighted bag-of-words vector. Figure 7.1 displays the cosine similarities between the vectors we obtained for the different communities. As can be seen, and somewhat unsurprisingly, there is a high degree of overlap between *migrants* and *immigrants*. For this reason, these two communities/keywords will be merged for the analyses in this chapter. We can furthermore see that *migrants* and *refugees* are also somewhat similar in the dataset, but since the similarity between *immigrants* and *refugees* is much lower, we keep *refugees* as a separate community. Note that we omitted the keyword *hopeless* in Figure 7.1, as we found this keyword to be too generic to be viewed as describing a particular community. For this reason, we will not consider this keyword in our community-specific experiments and analysis.

Community	Associated terms
Immigrants	First-generation, resentment, cultures, foreign-born, undocumented, sentiment, spouses, applicant, citizenship, sanctuary
Migrants	Hatred, incoming, dreamers, coast, trafficking, racism, protections, deported, gangs, rescued, hostile, threatened, tortured, smuggling
Refugees	Repatriation, offshore, queer, seekers, resettlement, camps, fled, abuses, mercy, forget, envoy, cap, asylum, intake, diplomat
In need	Donor, desperately, Christ, drought, kindness, foster, budgets, compassionate, humanitarian, blankets, gifts, salvation, foster
Poor families	Diapers, nutritious, scholarship, rice, poverty, expenses, savings, malnutrition, babies, orphans, unhealthy, talented, Medicaid, grants
Vulnerable	Droughts, prey, strategies, hub, resilience, crop, proactive, exploitation, fragile, hazards, defence, weapons, instability, debt, threats
Women	Feminist, maternity, abortions, husbands, beauty, fertility, unsafe, empowering, motivated, honour, equality, harassment, roles, ratio, attractive
Disabled	Assistive, pension, impaired, heroes, integrating, consideration, allowance, disadvantaged, begging, career, disabled-friendly, sacrificed
Homeless	Downpour, jobless, addicts, evicted, shelters, hungry, streets, rough, roofs, soup, sleepers, drop-in, devastated, tents, fortunate

Table 7.3: Selection of terms found for the different communities, with $k = 100$.

In Sections 7.4 and 7.5, we describe experiments in which PCL classifiers are trained in a way that (partially) prevents them from learning about community-specific thematic PCL. This will allow us to better characterise the abilities of fine-tuned language models, as the overlap between the themes covered by the training and test sets is reduced.

7.4 Omitting Community-Specific Training Data

Our main hypothesis, as outlined in the introduction of this chapter, is that the SemEval PCL detection task is easier than one might expect because it involves a combination of linguistic PCL, which is easier to detect, and thematic PCL. While we believe that thematic PCL can be hard to detect in general, our hypothesis is that

Figure 7.1: Similarity between the different communities from the DPM! dataset.

it is simplified, in the context of the SemEval task, because of the overlap between the themes covered in the training and test data of the DPM! dataset. If a language model is truly able to recognize PCL, then it should be capable of identifying (thematic) PCL about communities it has not seen during training. In this section, we report the results of an experiment where we test the performance of the model per community in two settings. First, we consider the standard setting, where the model has had access to the entire training set. Second, we consider the setting where all examples about the community being tested were removed from the training set. Note that for the latter case, we need to train a separate model for every community, each time omitting the corresponding training examples.

The results, which are reported in terms of F1 score $\%$ and averaged over 5 runs,

	Full Training	**Comm. Omitted**
Migr. + Imm.	**43.6** $_{7.89}$	25.3 $_{3.27}$
Refugees	50.4 $_{8.36}$	**54.0** $_{5.12}$
In need	**55.3** $_{3.12}$	51.2 $_{1.04}$
Poor families	52.7 $_{6.34}$	**53.7** $_{7.18}$
Vulnerable	**54.7** $_{3.75}$	51.6 $_{3.29}$
Women	31.5 $_{8.79}$	**41.7** $_{7.53}$
Disabled	**54.6** $_{5.52}$	52.4 $_{3.85}$
Homeless	**60.2** $_{1.85}$	54.4 $_{2.49}$
All communities	55.4 $_{0.46}$	-

Table 7.4: Performance of RoBERTa-base models fine-tuned with (Full Training) and without (Comm. Omitted) training examples from the testing community.

are summarized in Table 7.4, where we also report the standard deviation. We can make a number of clear observations. First, the performance of the model that was trained on the full training set varies substantially across the different communities. For instance, the F1 score for *homeless* is almost twice as high as that for *women*. Second, excluding training examples about the test community has a substantial impact on the results for some communities, but not for others. For *migrants + immigrants*, we can see a particularly large drop in performance, which suggests that PCL towards this community is more likely to be thematic than for the other communities. For some of the other communities, we also see drops, although these are much smaller. Surprisingly, for some communities, the performance improves when omitting training examples from that community, which is most pronounced for *women*. This suggests that PCL towards women is more likely to be linguistic (and thus community-independent), while the model may have learned incorrect associations from the themes that are present in the training examples about women. This will be further explored in the qualitative analysis in Section 7.6.

7.5 Masking Community-Specific Terms

We also conduct a variant of the experiment from the previous section. In this case, no training examples are removed, but we instead mask (some) occurrences of community-related terms, as identified in Section 7.3.2, in the training data. Note that we mask occurrences of such terms regardless of the community a training example is about (e.g. a term that was identified for *refugees* would still be masked in examples about *immigrants*). This setup has the advantage that the number of training examples remains constant. Moreover, the model may now also be prevented from learning thematic PCL by training on related communities. For instance, in the setting from Section 7.4, the model may be able to learn condescending themes about the *homeless* community from training examples mentioning the *vulnerable* keyword.

The results are reported in Tables 7.5, 7.6 and 7.7, where the masking probability for mentions of community-related terms is varied from 0% to 100%. We also report the standard deviations over 5 runs. In all tables, configurations which outperform the baseline (i.e. the setting where the original training set is used) are shown in bold, while the best overall result for each community is underlined. Results are reported in terms of F1 score (%) and are averaged over 5 runs. The main findings from Section 7.4 are confirmed by this experiment. In particular, for *migrants + immigrants*, we find that masking community-related terms leads to a substantial drop in performance (especially when 100% of the mentions are masked). This again suggests that the classifier, in the standard setting, heavily relies on the fact that condescending themes from the test set are also present in the training set. For *women*, we can see that masking can improve the results, which again suggests that the type of PCL for this community is mostly linguistic. In fact, for all but one community, namely *migrants + immigrants*, the best overall results are obtained with some degree of masking. This suggests that linguistic PCL is prevalent across the dataset, and that the fine-tuned RoBERTa-base model is susceptible to lean incorrect associations

	Top-25 community based terms										Baseline	
	100%		80%		60%		40%		20%		0%	
Migr. + Imm.	33.80	8.19	35.89	4.75	33.15	5.73	36.28	7.86	43.01	8.92	**43.6**	7.89
Refugees	49.15	4.14	46.63	6.15	**51.66**	3.87	**53.04**	4.25	**51.16**	3.73	50.4	8.36
In need	**56.49**	2.27	**55.93**	4.91	**56.85**	3.69	**57.38**	2.86	**56.12**	3.55	55.3	3.12
Poor families	**57.22**	2.21	51.90	4.98	47.57	6.57	**56.03**	1.76	**53.44**	5.54	52.7	6.34
Vulnerable	53.41	1.86	52.49	4.24	**56.62**	5.58	**55.72**	4.51	**55.71**	6.83	54.7	3.75
Women	**39.96**	6.07	**34.34**	7.93	**31.75**	9.37	**32.33**	9.73	31.47	8.63	31.5	8.79
Disabled	**56.38**	2.38	53.04	4.99	51.60	7.52	50.08	3.32	49.63	1.90	54.6	5.52
Homeless	57.78	2.32	54.29	2.14	55.80	5.95	57.24	3.08	55.10	4.73	**60.2**	1.85
All comm.	53.82	1.60	51.27	2.79	51.66	4.18	53.21	1.30	52.40	0.76	**55.4**	0.46

Table 7.5: Performance of RoBERTa-base models fine-tuned on variants of the training set in which community-related terms are masked. Results are shown with the $k = 25$ top terms from each community, and with varying masking probabilities.

.

	Top-100 community based terms										Baseline	
	100%		80%		60%		40%		20%		0%	
Migr. + Imm.	27.7	4.76	38.0	8.78	31.6	8.56	35.7	6.83	40.0	3.02	**43.6**	7.89
Refugees	49.9	3.17	50.1	6.11	47.1	2.81	**52.2**	3.76	**53.0**	3.61	50.4	8.36
In need	**55.6**	1.19	55.2	1.21	**55.8**	1.87	**56.5**	1.45	**58.6**	3.77	55.3	3.12
Poor families	**55.9**	3.31	**57.5**	3.05	52.0	6.93	47.8	6.24	52.7	4.89	52.7	6.34
Vulnerable	54.3	6.28	**56.8**	4.80	52.7	7.90	**57.5**	3.70	**55.8**	6.27	54.7	3.75
Women	31.0	9.92	**37.6**	5.44	**39.3**	4.91	**41.0**	2.74	**39.7**	3.97	31.5	8.79
Disabled	51.8	2.81	49.3	5.59	52.4	5.23	48.7	2.42	48.0	4.52	**54.6**	5.52
Homeless	58.5	0.79	58.4	2.94	57.8	2.64	57.7	5.22	**62.1**	1.95	60.2	1.85
All comm.	52.3	1.49	53.4	2.15	51.6	1.70	52.5	1.39	53.9	0.87	**55.4**	0.46

Table 7.6: Performance of RoBERTa-base models fine-tuned on variants of the training set in which community-related terms are masked. Results are shown with the $k = 100$ top terms from each community, and with varying masking probabilities.

.

between thematic terms and the presence of PCL.

7.6 Qualitative Analysis

The experiments in Sections 7.4 and 7.5 have revealed stark differences in the robustness of PCL detection models across different communities, when the model is (partially) prevented from learning community-specific themes during training. In

	Top-500 community based terms										Baseline	
	100%		80%		60%		40%		20%		0%	
Migr. + Imm.	25.2	6.82	34.3	6.89	42.3	6.47	36.0	7.24	34.9	6.77	**43.6**	7.89
Refugees	49.6	7.72	49.5	1.60	48.1	2.56	48.5	5.30	**53.5**	4.85	50.4	8.36
In need	**56.9**	1.10	54.7	1.65	**58.6**	2.44	**57.1**	2.80	55.1	3.46	55.3	3.12
Poor families	51.7	3.90	52.2	5.92	52.1	5.60	50.2	4.44	46.6	2.62	52.7	6.34
Vulnerable	48.4	3.33	47.5	2.26	**56.3**	6.12	54.1	5.35	52.3	2.47	54.7	3.75
Women	**38.2**	2.60	**39.8**	6.05	**39.9**	8.62	**39.5**	4.76	**35.9**	7.10	31.5	8.79
Disabled	45.8	3.15	46.3	2.06	54.4	4.43	52.1	6.51	53.0	4.54	**54.6**	5.52
Homeless	54.6	1.86	54.9	2.63	**61.3**	3.01	60.0	1.84	57.9	5.74	60.2	1.85
All communities	51.2	3.59	50.7	1.15	54.6	2.59	52.9	2.59	52.3	1.97	**55.4**	0.46

Table 7.7: Performance of RoBERTa-base models fine-tuned on variants of the training set in which community-related terms are masked. Results are shown with the $k = 500$ top terms from each community, and with varying masking probabilities.

particular, our results suggest that PCL examples for *migrants + immigrants* are often thematic in nature, with the same themes recurring in both the training and test sets. Conversely, the results for *women* suggest that PCL towards that community is more likely to be linguistic in nature. In this section, we supplement our findings with a qualitative analysis, where we focus on these two communities.

Migrants + Immigrants. In Table 7.8, we can see examples of PCL which were consistently[2] classified correctly when including the community in the training set, but where the model was unable to recognise the PCL when trained without examples from the test community. In bold, we highlight some community-specific themes that are common in examples of PCL and which the model should be unable to learn when not presented with similar examples during training. Note, for instance, that the word *Dreamer* is present in all the examples from this table. It thus seems safe to infer that the model has learned that this term, when associated to the *migrants + immigrants* community, is highly predictive of the presence of PCL, when such examples are included in the training data. The use of other terms

[2]We focus on cases where the classification is consistent across different runs of our experiments, i.e. with different random seeds, to reduce the influence of instances that were classified correctly or incorrectly by chance.

Classified correctly only with full training set
On the campaign trail, **Trump** promised to **deport** all **undocumented** migrants. Since taking office, he appeared to soften on **dreamers**, a relatively well-educated and industrious group who he described as "incredible kids"
But without resolution, the centrists warn they will have enough petition signatures by Tuesday to force House votes later this month, including on their preferred bill which provides young **"Dreamer"** immigrants **protection** from **deportation** and a chance to apply for **citizenship**.
Passage of the measure came over the opposition of Democratic leaders who demanded the promise of a vote to protect **"Dreamer"** immigrants brought to the country illegally as children. A band of tea party Republicans was also against the legislation over what it sees as spiralling spending levels.
The New York senator said he was hopeful about talks on so-called **Dreamers**, more than 700,000 young immigrants brought to the US as children who were **protected** under the Obama-era Deferred Action for Childhood Arrivals **(Daca)** programme.

Table 7.8: Examples of PCL for *migrants* + *immigrants*, which are consistently classified correctly when the model is trained on the full training set, but consistently misclassified when training examples from this community are excluded from the training set.

.

such as *deportation*, *undocumented* or *citizenship* are also strongly related to the community and might help the model to identify the presence of PCL.

In contrast, the examples of PCL in Table 7.9 were consistently identified correctly, whether the training examples for *migrant* + *immigrant* were included or not. As expected, we can indeed think of these examples as being primarily *linguistic PCL*, in the sense that what makes them condescending is *how* the message is expressed, more than *what* is being expressed. In bold, we highlight the presence of some common linguistic features of PCL. For instance, in the first example, we can see an excess of flowery wording and adjectives to express a message, the use of metaphors and an almost poetic style to describe a vulnerable situation, which are common features of PCL [125]. The second and third examples also show clear differences in power and privilege, for instance, through the use of expressions such as *we have a moral responsibility*, *show them solidarity* or *permitting them to work and*

Classified correctly even without community-specific training examples
The Irish famine led to a massive influx of Irish immigrants to New York during the late 1840s and 1850s. As the **downtrodden** Irish escaped the famine in their home country, however, **they came to a place where life was just as tough**. Disembarking from **coffin ships**, Irish newcomers were **greeted with a new life of hardship, slums and tough, endless labor**.
Vatican City: As record numbers of people flee conflict, persecution and poverty, governments, citizens and the Church **have a moral obligation** to safeguard migrants and **show solidarity** with them, the Pope has said.
Barack Obama implemented the DACA program five years ago to **help bring** the children of undocumented immigrants **out of the shadows** of illegality, **permitting them** to study and work **without fear**.
It's been hard **breaking through the barrier of migrant communities**. Many women from my own community do not take my work seriously and do not support it, and **I grapple with this**. **I'm trying to help, to make things better, but many women find comfort in the norms and the way things are**.

Table 7.9: Examples of PCL for *migrants* + *immigrants*, which are consistently classified correctly both when including or excluding the community from the training set.

study without fear. The last example conveys a distance between the author and the community (*breaking through the barrier of migrant communities*) and expresses presuppositions and an authority voice based on the idea of a *saviour-victim* relation (*I grapple with this*, *I'm trying to help, to make things better, but many women find comfort in the norms and the way things are*). These examples of *linguistic PCL* are independent of the community they are addressing, which is why the model still recognises them even when no training examples for the *migrants* + *immigrants* community are provided.

Women. Table 7.10 shows examples of PCL that were missed when using the full training set, but consistently classified correctly when omitting *women* examples. In the first paragraph, the phrase *their shame continues*, a community-independent value judgement, makes the text condescending. The second and third example express a *saviour-victim* relation, where the differences between power and vulner-

Classified correctly only without community-specific training examples
Many of these women now lie in unmarked graves, a situation that is slowly being rectified by the work of the voluntary Justice for Magdalenes Group. **Their shame continues**.
However, "when **a major male rock star who could do anything at all with his life decides to focus on the rights of women and girls worldwide** - well, all **that's worth celebrating**. We're proud to name that rock star, Bono, our first Man of the Year," it said.
A Cosmopolitan spokesperson says with a focus on empowerment, **the magazine is "proud of all that the brand has achieved for women around the world"**.

Table 7.10: Examples of PCL for *women*, which are classified correctly only when excluding the community from the training set.

ability, as well as an admiration towards the *saviour*, are explicitly stated. As these examples are clearly linguistic, we can expect that a model which has not seen *women* examples should be able to classify them correctly. Surprisingly, all three paragraphs were missed by the model that was trained on the full training data. To understand why this is the case, note that 95% of the training examples for *women* are negative. As a result, several of the terms that are associated with women (almost) exclusively appear in negative training examples. This can lead the model to believe that these words are indicative of a lack of PCL. By masking community-related terms, or omitting training examples from this community entirely, we can prevent the model from learning such coincidental associations. In the table, we highlight in bold the presence of some common linguistic features of PCL.

Other communities. For illustrative purposes we also include Table 7.11, where we show examples of instances which were correctly classified only when some degree of stereotypes masking was applied, but were missed when training either on the full data or on the data which omitted the testing community. These results are expected in the sense that the model still needs to learn about community-specific themes to identify thematic PCL. However, some degree of masking can prevent the model from learning wrong associations between stereotypes and PCL, as seems to

Classified correctly only with partial masking
"Eleven months into his administration, the country is showing signs of progress in most sectors of the economy. With the implementation of the free senior high school programme, most students, **especially those from poor families**, **who hitherto would not have progressed to the senior high school**, **have the opportunity now to receive** secondary education **to make them better and more functional in society**", Dr Nyarko said.
Today, Brooklyn is home to people of all races, most struggling to make ends meet. Council flats continue to degrade as the population swells – **unemployment and homelessness sees people of different races** lining up side-by-side **for a plate of free food**. It's a representation of **the rainbow nation in trauma**, **with its colours dulled and blended together by suffering**.
Helping refugee children fit in a bonus for Juventus football camp.
Swimming superstar Adam Peaty is set to unveil **a new motorbike for charity** in memory of schoolgirl Imogen Evans, who used the service. The Shropshire and Staffordshire Blood Bikes is a charity which **saves lives** by delivering vital blood supplies to those in need.
RADIO Veritas, the leading faith-based AM station in Mega Manila, continues **its commitment to charity and public service** through an initiative dubbed as "**Good Samaritan**". Since it was launched last June 2017 (airing every Monday to Friday from 1-2 p.m.), Radio Veritas has listed 182 cases of pleads and requests that have been fulfilled through this program. It serves as a platform **for those in need to make on-air appeals** for legal, spiritual, medical, material and financial assistance, and **link them to "Good Samaritans" who are willing to share**.

Table 7.11: Examples of PCL for different communities which are consistently classified correctly when partially masking community-related terms, but that are missed when training either on all data or removing all the community-specific training examples.

happen with the data for *women*. Tentatively, partially removing themes and stereotypes might force the model to learn more about linguistic features of condescending language, or even to make some commonsense lead decisions, as it can not fully rely on previously seen stereotypes. Nevertheless, further research on this direction will be needed to confirm this hypothesis. In the table, we highlight in bold both, common linguistic features of PCL and community-specific themes or stereotypes found in DPM!

7.7 Summary

In this chapter we have analyzed the the DPM! dataset and its influence on the challenge of detecting Patronizing and Condescending Language, with the aim of improving our understanding of the nature of such of language. We highlighted the distinction between two types of PCL. On the one hand, linguistic PCL is concerned with how the message is expressed and is largely community-independent. On the other hand, thematic PCL is more concerned with the message itself, and often relates to aspects that are highly community-specific. Our analysis suggests that for some communities, instances of PCL are mostly linguistic, while for other communities, thematic PCL is more prevalent. Moreover, detecting thematic PCL remains highly challenging in settings where the training data does not include examples covering similar themes. A better understanding of these phenomena can help future work to improve the detection of PCL and, eventually, contribute to more responsible and inclusive communication. As a first step, we envisage that a more fine-grained annotation of PCL detection datasets will be needed, distinguishing between (subcategories of) linguistic and thematic PCL, to help us train better models and allow for a more insightful evaluation.

Chapter 8

Conclusions and Future Work

8.1 Introduction

In this last chapter we conclude the thesis by summarizing our contributions and main findings. We revisit the motivation to conduct this research, as well as the hypothesis that inspired our work. We relate how our main findings have answered the research questions which were posed at the beginning of this journey and suggest some lines of research for future work.

8.2 Thesis Summary and Contributions

This thesis emerges from observing the treatment that vulnerable communities often receive in the media, and the motivation to make steps towards more responsible and inclusive communication.

The relationship between language and power has been extensively studied throughout history. As one of the two sides of unbalanced power relations, privileged communities use language to describe those more underrepresented groups with different types of discourse. *Hate speech*, *offensive language* or *rumour spreading* are commonly directed towards underrepresented communities. But there is another type of language, namely PCL, which while in a more subtle way and often with

good intentions, can be equally harmful, especially when widespread by the media. The discourse of condescension undermines vulnerable communities, feeds stereotypes and supports power relations, reinforcing the dichotomy between *saviours* and *victims* [104, 116, 51, 52, 9, 162]. PCL underestimates deep-rooted societal problems, suggests shallow and ephemeral solutions, and fuels discriminatory behaviour towards those receiving the condescending treatment [115].

While other, more flagrant types of harmful language have been in the focus of NLP research for many years, we found that PCL had been mainly neglected by the field. Although some previous works tried to address either condescending messages in social media [175] or closely related discourses [151, 104], the study of (unintended) PCL towards vulnerable communities in the media remained an unexplored area.

The detection of this gap, the identification of the challenges that PCL presented, and the potential positive impact of our research, motivated this thesis.

With the objective of sharing our conclusions, we revisit our two-fold hypothesis, which was introduced in Chapter 1 and reads as follows:

> 1) Patronizing and Condescending Language can, to some extent, be automatically detected and categorized by Language Models, in spite of its subtle and subjective nature. However, its detection often requires commonsense reasoning, as well as world knowledge and an understanding of human values, which will pose a challenge for NLP models.
>
> 2) The analysis of PCL from an NLP perspective can help us to improve our understanding of the nature and features of PCL towards vulnerable communities.

After concluding and presenting our work, we can say that our hypothesis is confirmed. The works summarized in this thesis support this statement.

Chapter 2 served to put our research in context, reviewing the most relevant previous works which had been developed in the study of condescension and closely related

topics, either in NLP or in other areas. In Chapter 3 we presented Don't Patronize Me!, a novel dataset annotated with PCL towards vulnerable communities created to address the lack of specific data. With the objective of delving into the study of PCL, we also created a taxonomy of PCL categories used by media sources towards vulnerable communities. After a careful curation of the data, up to five trained annotators participated in a two-level annotation process. First, they assessed a paragraph as containing or not PCL, or being a borderline case. Then, they took all agreed positive cases of condescension to identify which categories of PCL were present in the paragraph. The process of curation and annotation of DPM! already allowed us to identify features and challenges of such language, such as its subtlety and subjectivity. Moreover, we identified that world knowledge was often required to detect PCL and that personal values could play an important role in classifying a message as condescending. These findings, which started to partially confirm our hypothesis, arose as challenges that our research would try to overcome. DPM! has been the seed data for our research, but it is also intended to be a new linguistic resource for the community.

Chapter 4, which contains the first baseline results, showed that Language Models were indeed able to identify and categorize PCL to some extent. However, the results also left a large margin for improvement. A closer, qualitative look at some examples correctly or wrongly classified by the models reinforced the intuition that PCL detection is a challenging task which might need specific approaches to be solved.

In Chapter 5 we explored what kind of previous knowledge could help a model to better identify PCL. We also experimented with different pre-training strategies to find which one is more suitable for this task. Specifically, we used full fine-tuning of a pre-trained RoBERTa-based LM, and the use of adapters. In both cases, we trained the models on ten auxiliary tasks, whose data covered human values, flagrant forms of harmful language, political discourse and other *a priori* less related tasks, such

as sentiment analysis or irony detection. The pre-trained model was then fine-tuned on PCL detection. Our experiments showed that some gains are indeed possible by infusing the model with previous, specific knowledge, especially about offensive language, commonsense morality and hate speech. More surprisingly, data annotated with irony and sentiment also helped the model. Therefore, these experiments confirmed that PCL detection could benefit from previously learning about human values and other more or less related types of discourses. Results also pointed towards different auxiliary tasks helping the model to identify PCL of different categories, even when it is not always able to tell what category is present in the text. The qualitative analysis of the paragraphs which were missed by the baseline model (i.e., without pre-training in any auxiliary task) but correctly classified by each one of the best performing models (i.e., pre-trained on different related data), helped us to better understand the nature of PCL, the background needed to understand it and the challenges that NLP models still face solving this task.

Chapter 6 supported the hypothesis that LM can indeed identify and classify PCL to some extent, which we had already confirmed with experiments in Chapter 4. In fact, the best performing systems participating in the shared task in PCL detection hosted at SemEval-2022, showed that a judicious use of SoTA text classification techniques, such as data augmentation, introducing a contrastive loss function, ensembling different models or applying prompt learning, to name a few, could lead LMs to obtain results almost on pair with human performance. As much as we learnt from the organization and outcomes of this task, it also arose new questions which challenged the hypothesis that PCL detection was a subtle, difficult task which often required commonsense reasoning and world knowledge.

This situation led us to develop the experiments summarized in our 7th and last chapter, where we analyzed in depth the DPM! dataset. We unveiled that there are two types of PCL: *linguistic* PCL, where the condescension is expressed by how we formulate the message, and which is based mainly on community-independent

linguistic features; and *thematic* PCL, where the condescension comes with what is said, or the message itself. Thematic PCL is based mainly on themes or stereotypes closely related to the communities which are the target of the condescension. The overlap between the communities covered in the training and the test set allows the model to learn these condescending stereotypes and themes. The model would not be able to make such associations if those same stereotypes and themes were not recurrent in the training set. For instance, the model would not be able to identify (thematic) PCL towards unseen communities or if the themes associated with a specific community change over time. The experiments made in this direction show that there are indeed (at least) two types of PCL and that the task of PCL detection, as it is formulated, is easier to solve due to the overlap in the training and test communities.

In conclusion, every experiment developed in the framework of this thesis helped us confirm our two-fold hypothesis: We have shown how NLP models perform in PCL detection and what knowledge is helpful for NLP models to better identify PCL. In addition, the research conducted in the framework of this thesis has helped us to better understand the nature of PCL.

8.3 Research Questions and Main Findings

In this section, we revisit and discuss our research questions, introduced in Section 1.2, and try to answer them in relation to the outcomes and findings derived from our research.

Research Question 1. How easy it is for human annotators to identify PCL towards vulnerable communities? Do human annotators agree in their assessments about the presence of this type of language?

The literature review about condescension pointed us towards the notion of a subjective discourse which was difficult to identify even for humans, especially when directed towards vulnerable communities and with good intentions. The development of Chapter 3, with the definition of the task of PCL detection, the curation of the data and, especially, the annotation process, proved that PCL detection and classification poses a challenge not only for NLP models but also for humans. The task of identifying if a paragraph contained PCL turned out to be more difficult than identifying which categories of PCL were present in the text. This points towards the conclusion that, although the categories are well defined, which makes it easier for annotators to identify them, assessing if that category is expressing condescension is still subtle and subjective and a clear challenge even for humans. This justifies that the agreement between annotators was lower for subtask 1 (PCL detection) than for subtask 2 (PCL categorization).

Research Question 2. *To what extent can Language Models identify and categorize PCL? Which NLP techniques are best suited to address this challenge?*

Our findings in Chapter 4 reveal that, as expected, Language Models perform better in PCL detection than other architectures, such as models based on SVMs or Bi-LSTMs. Among the LMs that we tested, the best results are obtained by BERT-large, with RoBERTa-base and BERT-base yielding similar results. RoBERTa-large achieved a highly inconsistent performance. For the task of categorizing PCL, RoBERTa-large obtained the best results, both when omitting and including negative examples of PCL. However, the results by category are highly dependent on the number of training examples. Although LMs can indeed identify and categorize PCL, a *vanilla* approach (i.e., a model not customized for the specific task), leaves much room for improvement, especially for the less populated categories. The findings of Chapter 6 show us that standard text classification techniques are very useful for PCL detection. Specifically, balancing the class imbalance, applying a contrastive loss function while training, ensembling several of the best-performing models or ap-

plying prompt learning, can boost the results of both subtasks, namely binary and multi-label classification of PCL.

Research Question 3. *What does a model need to know to better identify PCL? To what extent would it need to understand human values?*

Although Chapter 5 was devoted to answering this question, all the findings we obtained through the thesis help us to better understand the nature of PCL in this regard, i.e., to understand how important it would be, for a model, to know about human values, or to contain world knowledge, to address the challenge of PCL detection. From the findings in Chapter 3 and 4, we conclude that PCL is hard to detect both for NLP models and for humans, hence the difficulties the annotators found in the annotation process. In Chapter 5 we experimented by infusing the model with knowledge from different areas, which we considered to be more or less related to PCL. We saw that gains are indeed possible when pre-training in some auxiliary tasks, especially those related to other types of harmful language or tasks where human values were addressed. Other more distant tasks, such as sentiment or irony detection also proved useful for our main challenge. A closer look allowed us to see that each task could help the model identify PCL belonging to different categories. Although the improvements in performance were humble, these results motivate further research in this area. Some ideas to continue exploring this line of research are presented in Section 8.4. Although the findings of Chapter 6 questioned the need for world knowledge and human values understanding, the experiments developed in the framework of Chapter 7 clarified this issue. The results obtained point to the fact that, although linguistic PCL might be easily learnt from training examples, identifying thematic PCL would require of more abstract understanding of the world and human values, if the themes or stereotypes it refers to are not included in the training data.

Research Question 4. Can current State-of-The-Art NLP models effectively generalize to address the complexity of PCL?

Chapter 3 presented PCL as a complex, subtle and subjective kind of language. The DPM! dataset in particular showed that annotating PCL was challenging, that context about the paragraph was often required and that the personal profile of the annotator could play an important role in considering a message as being condescending. With this picture of PCL, a good performance of NLP models in this task seemed unlikely, but the findings in Chapter 4 and especially in Chapter 6 showed a very different reality. SoTA techniques in NLP, combined with standard models, can actually get results almost on par with human performance, at least in a binary classification task (i.e., classifying a paragraph as a positive or negative example of PCL). However, the findings in Chapter 7 unveiled that the task, as it has been formulated, might be easier to address than the actual challenge of detecting PCL, due to the overlap between communities in the training and test set in DPM!. Therefore, we consider that PCL detection and categorization might require different approaches to address specific aspects of it.

8.4 Future Work

In this section, we offer some ideas for further research in PCL detection and categorization:

The data. The DPM! dataset is meant to be a useful resource for the study of PCL, but the data available in this research area is still scarce. For future datasets, we consider that it would be useful to cover a higher number of communities. In addition, we recommend that a different way to extract the data is applied, in order to identify texts referring to a given community even if a specific keyword is not mentioned. Moreover, the field could benefit from exploring different sources where

vulnerable communities might receive condescending treatment, such as social media messages, social responsibility reports, or fundraising campaigns.

Regarding the annotation, given that PCL is subtle and subjective, we would recommend training the annotation team thoroughly in the task, and counting on as many annotators as possible. Besides binary and multi-label (i.e., categories) annotation, annotating the messages with PCL as being mainly linguistic or mainly thematic might help the model to identify these two PCL phenomena.

As PCL is presented as a complex discourse, different approaches might be needed to address different aspects of PCL. More (and more finely-grained annotated) data would help explore different models for different categories or PCL, for instance, or for identifying linguistic vs thematic PCL.

Incorporating additional external knowledge. The experiments in Chapter 5 show that the performance of a model can improve if it is pre-trained in the proper data. Further research in this direction would include identifying potential related auxiliary tasks and finding associated data, ideally more than one dataset for each task. For instance, if stereotypes are considered to play an important role in (thematic) PCL, multiple datasets on stereotypes should be compared. Also, combining several auxiliary tasks could benefit the performance of the model. Further research could also focus on exploring which previous knowledge and to what extent benefits the model when targeting the detection of specific categories of PCL.

Exploring PCL towards non-vulnerable communities. PCL is not exclusive of vulnerable communities but, to the best of our knowledge, there is no research that compares the features of condescending language targeting different communities. It would be interesting to test to what extent a model trained on PCL towards vulnerable communities can identify PCL towards other groups which, while not being vulnerable, are susceptible to receive this kind of treatment, such as vegans, young

people or people with strong religious beliefs. Further analysis in this direction would help us better understand the nature of PCL. In addition, building a model able to learn PCL towards a broader spectrum of communities would potentially improve a system for flagging condescending language.

Multidisciplinary research. The study of PCL from an NLP perspective will benefit from insights from other disciplines, such as Psychology, Sociology or Cultural, Media and Political Studies, to name a few. Contributions from these areas would have been extremely useful to better understand the nature of PCL and the results we were obtaining with our experiments. Furthermore, a study about PCL towards vulnerable communities should count on the perspective of these groups. Therefore, counting on representatives of the communities being studied would contribute valuable insights and will make the research process fairer and more responsible.

8.5 Final Remarks

With the works developed in the framework of this thesis, we have aimed at introducing the detection and classification of PCL towards vulnerable communities as a new research topic in NLP.

This research is born with the objective of contributing to more ethical communication which could, in turn, help to build a more equal society. Crucially, the use of PCL towards vulnerable communities in the media is often unintentional, hence developing tools that flag instances of PCL, which could work similarly to spelling and grammar checkers, can bring about meaningful change. This makes PCL detection an important social challenge that should be addressed by the NLP community.

Last, we would like to highlight that by doing this research, we might have been condescending ourselves, as we have aimed to detect condescension towards com-

munities we are not part of, without asking their members about what they might consider condescending. This is another proof that PCL is unconscious and maybe even difficult to avoid. However we still think that our work is useful as a way to keep giving steps towards more responsible communication, where everyone could feel they are represented in a fair way.

Bibliography

[1] K. Adaikkan and T. Durairaj. Ssn_nlp_mlrg at semeval-2022 task 4: Ensemble learning strategies to detect patronizing and condescending language. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 400–404, 2022.

[2] S. Agrawal and R. Mamidi. Lastresort at semeval-2022 task 4: Towards patronizing and condescending language detection using pre-trained transformer based models ensembles. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 352–356, 2022.

[3] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams. “the enemy among us” detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, 13(3):1–26, 2019.

[4] M. Banko, B. MacKeen, and L. Ray. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.16. URL `https://www.aclweb.org/anthology/2020.alw-1.16`.

[5] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020.

[6] J. Barnes, L. A. M. Oberländer, E. Troiano, A. Kutuzov, J. Buchmann, R. Agerri, L. Øvrelid, and E. Velldal. Semeval-2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Seattle. Association for Computational Linguistics*, 2022.

[7] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL `https://www.aclweb.org/anthology/S19-2007`.

[8] C. Basta, M. R. Costa-Jussà, and N. Casas. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384, 2021.

[9] K. M. Bell. Raising Africa?: Celebrity and the rhetoric of the white saviour. *PORTAL Journal of Multidisciplinary International Studies*, 10(1), 2013.

[10] E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[11] Y. Bestgen. Satlab at semeval-2022 task 4: Trying to detect patronizing and condescending language with only character and word n-grams. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 490–495, 2022.

[12] S. Bhatt and M. Shrivastava. Tesla at semeval-2022 task 4: Patronizing and condescending language detection using transformer-based models with data augmentation. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 394–399, 2022.

[13] E. Bodner. On the origins of ageism among older and younger adults. *International Psychogeriatrics*, 21(6):1003–1014, 2009.

[14] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[15] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, and F. D'Errico. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing & Management*, 60(1):103118, 2023.

[16] P. Bourdieu. *Language and symbolic power*. Harvard University Press, 1991.

[17] L. L. Brons. Othering, an analysis. *Transcience, a Journal of Global Studies*, 6(1), 2015.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

[19] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

[20] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[21] P. Burnap and M. L. Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15, 2016.

[22] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[23] M. K. Canales. Othering: Toward an understanding of difference. *Advances in Nursing Science*, 22(4):16–31, 2000.

[24] D. Caspi and N. Elias. Don't patronize me: media-by and media-for minorities. *Ethnic and Racial Studies*, 34(1):62–82, 2011.

[25] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2005. URL `https://www.aclweb.org/anthology/S19-2005`.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[27] J. Chhillar. Taygete at semeval-2022 task 4: Roberta based models for detecting patronising and condescending language. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 496–502, 2022.

[28] L. Chouliaraki. *The spectatorship of suffering*. Sage, 2006.

[29] L. Chouliaraki. Post-humanitarianism : Humanitarian communication beyond a politics of pity. *International Journal of Cultural Studies*, 2010.

[30] N. J. Conroy, V. L. Rubin, and Y. Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[31] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

[32] G. Da San Martino, S. Yu, A. Barrón-Cedeno, R. Petrov, and P. Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5640–5650, 2019.

[33] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.semeval-1.186`.

[34] M. Davies. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. available online at https://www.english-corpora.org/now/, 2013.

[35] Y. Deng, C. Dou, L.-Y. Chen, D. Miao, X. Sun, B. Ma, and X. Li. Beike nlp at semeval-2022 task 4: Prompt-based paragraph classification for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 319–323, 2022.

[36] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, 2017.

[37] F. Dervin. Cultural identity, representation and othering. In *The Routledge handbook of language and intercultural communication*, pages 195–208. Routledge, 2012.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

[39] L. Díaz-Rico. Tools for discourse analysis. In *The Immigration & Education Nexus*, pages 149–159. Springer, 2012.

[40] E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.656. URL https://www.aclweb.org/anthology/2020.emnlp-main.656.

[41] H. Dong, V. Suárez-Paniagua, W. Whiteley, and H. Wu. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728, 2021.

[42] P. Draper. Patronizing speech to older patients: A literature review. *Reviews in Clinical Gerontology*, 15(3-4):273–279, 2005.

[43] A. Edalat, Y. Yaghoobzadeh, and B. Bahrak. Aliedalat at semeval-2022 task 4: Patronizing and condescending language detection using fine-tuned language models, bert+ bigru, and ensemble models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 387–393, 2022.

[44] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[45] N. Fairclough. *Language and power*. Routledge, 2013.

[46] N. Fairclough and L. Chouliaraki. *Discourse in late modernity.* Edinburgh University Press, 1999.

[47] J. Fattahi and M. Mejri. Spaml: a bimodal ensemble learning spam detector based on nlp techniques. In *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)*, pages 107–112. IEEE, 2021.

[48] D. M. Felblinger. Bullying, incivility, and disruptive behaviors in the healthcare setting: identification, impact, and intervention. *Frontiers of health services management*, 25(4):13, 2009.

[49] Z. Feng, J. Tang, J. Liu, W. Yin, S. Feng, Y. Sun, and L. Chen. Alpha at semeval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104, 2021.

[50] J. R. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[51] S. T. Fiske. Controlling other people: The impact of power on stereotyping. *American psychologist*, 48(6):621, 1993.

[52] M. Foucault. *Power/knowledge: Selected interviews and other writings, 1972-1977*. Vintage, 1980.

[53] M. Foucault. *Archaeology of knowledge*. routledge, 2013.

[54] R. Fowler, B. Hodge, G. Kress, and T. Trew. *Language and control*. Routledge, 2018.

[55] B. Gaind, V. Syal, and S. Padgalwar. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*, 2019.

[56] B. George, S. Adarsh, N. Prajapati, B. Premjith, and S. Kp. Amrita_cen at semeval-2022 task 4: Oversampling-based machine learning approach for

detecting patronizing and condescending language. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 515–518, 2022.

[57] S. Gilda, L. Giovanini, M. Silva, and D. Oliveira. Predicting different types of subtle toxicity in unhealthy online conversations. *Procedia Computer Science*, 198:360–366, 2022.

[58] H. Giles and A. Williams. Patronizing the young: Forms and evaluations. *The International Journal of Aging and Human Development*, 39(1):33–53, 1994.

[59] H. Giles, S. Fox, and E. Smith. Patronizing the elderly: Intergenerational evaluations. *Research on Language and Social Interaction*, 26(2):129–149, 1993.

[60] H. Giles, Y. Zwang-Weissman, and C. Hajek. Patronizing and policing elderly people. *Psychological Reports*, 95(3):754–756, 2004.

[61] Y. Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[62] H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy, Aug. 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W19-3621`.

[63] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[64] G. Gorrell, M. E. Bakir, M. A. Greenwood, I. Roberts, and K. Bontcheva. Race and religion in online abuse towards uk politicians. *arXiv preprint arXiv:1910.00920*, 2019.

[65] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski. Semeval-2019 task 7: Rumoureval, determining rumour vera-

city and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, 2019.

[66] L. Gui, R. Xu, Q. Lu, J. Du, and Y. Zhou. Negative transfer detection in transductive transfer learning. *International Journal of Machine Learning and Cybernetics*, 9(2):185–197, 2018.

[67] J. Habermas. *The theory of communicative action: Volume 1: Reason and the rationalization of society*, volume 1. Beacon press, 1985.

[68] M. Hall, L. van der Maaten, L. Gustafson, and A. Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

[69] X. Han and Y. Tsvetkov. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.622. URL https://www.aclweb.org/anthology/2020.emnlp-main.622.

[70] M. Hashemi and H. Karimi. Weighted machine learning. *Statistics, Optimization and Information Computing*, 6(4):497–525, 2018.

[71] P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2020.

[72] B. W. Head et al. Wicked problems in public policy. *Public policy*, 3(2):101, 2008.

[73] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[74] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[75] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[76] D. Hovy and S. Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

[77] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

[78] D. Hu, Z. Mengyuan, X. Du, M. Yuan, J. Zhi, L. Jiang, M. Yang, and X. Shi. Pali-nlp at semeval-2022 task 4: Discriminative fine-tuning of transformers for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 335–343, 2022.

[79] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

[80] T. Huckin. Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, 155:176, 2002.

[81] T. Huckin. Textual silence and the discourse of homelessness. *Discourse & Society*, 13(3):347–372, 2002.

[82] T. Huckin, J. Andrus, and J. Clary-Lemon. Critical discourse analysis and rhetoric and composition. *College composition and communication*, pages 107–129, 2012.

[83] M. L. Hummert and D. C. Mazloff. Older adults' responses to patronizing advice: Balancing politeness and identity in context. *Journal of Language and Social Psychology*, 20(1-2):168–196, 2001.

[84] A. Iyer and S. Vosoughi. Style change detection using bert. In *CLEF (Working Notes)*, 2020.

[85] Z. Jin, S. Levine, F. Gonzalez Adauto, O. Kamal, M. Sap, M. Sachan, R. Mihalcea, J. Tenenbaum, and B. Schölkopf. When to make exceptions: Exploring language models as accounts of human moral judgment. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28458–28473. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/b654d6150630a5ba5df7a55621390daf-Paper-Conference.pdf`.

[86] M. Johnson. How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11, 2009.

[87] M. Kanakaraj and R. M. R. Guddeti. Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)*, pages 169–170. IEEE, 2015.

[88] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, 2019.

[89] M. Klemen and M. Robnik-Šikonja. Ulfri at semeval-2022 task 4: Leveraging uncertainty and additional knowledge for patronizing and condescending lan-

guage detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 525–532, 2022.

[90] R. H. Knapp. A psychology of rumor. *Public opinion quarterly*, 8(1):22–37, 1944.

[91] D. Koleczek, A. Scarlatos, P. L. Pereira, and S. M. Karkare. Umass pcl at semeval-2022 task 4: Pre-trained language model ensembles for detecting patronizing and condescending language. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 445–453, 2022.

[92] M. S. Komrad. A defence of medical paternalism: maximising patients' autonomy. *Journal of medical ethics*, 9(1):38–44, 1983.

[93] J. Lacan. *The Ego in Freud's Theory and in the Technique of Psychoanalysis, 1954-1955*, volume 2. WW Norton & Company, 1991.

[94] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[95] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643, 2022.

[96] E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. What values should an agent align with? *Autonomous agents and multi-agent systems*, 36(1):1–32, 2022.

[97] Q. Liu, M. J. Kusner, and P. Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

[98] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[99] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of NAACL-HLT*, pages 615–621, 2019.

[100] H. H. Mao. A survey on self-supervised pre-training for sequential transfer learning in neural networks. *arXiv preprint arXiv:2007.00800*, 2020.

[101] B. D. Margić. Communication courtesy or condescension? linguistic accommodation of native to non-native speakers of english. *Journal of English as a lingua franca*, 6(1):29–55, 2017.

[102] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[103] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[104] J. Mendelsohn, Y. Tsvetkov, and D. Jurafsky. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:55, 2020.

[105] D. L. Merskin. *Media, minorities, and meaning: A critical introduction*. Peter Lang, 2011.

[106] S. Meyer, M. Schmidhuber, and U. Kruschwitz. Ms@ iw at semeval-2022 task 4: Patronising and condescending language detection with synthetically generated data. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 363–368, 2022.

[107] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[108] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[109] S. Mohammad, E. Shutova, and P. Turney. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, 2016.

[110] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

[111] A. Mosquera. Amsqr at semeval-2022 task 4: Towards autonlp via meta-learning and adversarial data augmentation for pcl detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 485–489, 2022.

[112] M. Nadeem, A. Bethke, and S. Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, 2021.

[113] P. Nandwani and R. Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):1–19, 2021.

[114] J. Nathanson. The pornography of poverty: Reframing the discourse of international aid's representations of starving children. *Canadian Journal of Communication*, 38(1), 2013.

[115] S. H. Ng. Language-based discrimination: Blatant and subtle forms. *Journal of Language and Social Psychology*, 26(2):106–122, 2007.

[116] D. Nolan and A. Mikami. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. *Global Media and Communication*, 9(1):53–70, 2013.

[117] D. Nozza, E. Fersini, and E. Messina. Deep learning and ensemble methods for domain adaptation. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 184–189. IEEE, 2016.

[118] J. Oldenburg, J. Aparicio, J. Beyer, G. Cohn-Cedermark, M. Cullen, T. Gilligan, U. De Giorgi, M. De Santis, R. de Wit, S. Fosså, et al. Personalizing, not patronizing: the case for patient autonomy by unbiased presentation of management options in stage i testicular cancer. *Annals of Oncology*, 26(5): 833–838, 2015.

[119] A. Oliver. The'pornography of poverty'and the'brothel without walls': Understanding the impact of art on development. *Undercurrent*, 3(2), 2006.

[120] T. O'reilly. *What is web 2.0*. " O'Reilly Media, Inc.", 2009.

[121] F. Ortiz. In-domain and cross-domain classification of patronizing and condescending language in social media and news texts: A study in implicitly aggressive language detection and methods. Master's thesis, Uppsala University, 2022.

[122] P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. PYKL, B. Gambäck, T. Chakraborty, T. Solorio, and A. Das. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.semeval-1.100`.

[123] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[124] C. Perez-Almendros and S. Schockaert. Identifying condescending language: A tale of two distinct phenomena? In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 130–141, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlp4pi-1.15.

[125] C. Perez-Almendros, L. Espinosa Anke, and S. Schockaert. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.518. URL https://www.aclweb.org/anthology/2020.coling-main.518.

[126] C. Perez-Almendros, L. Espinosa-Anke, and S. Schockaert. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.38. URL https://aclanthology.org/2022.semeval-1.38.

[127] C. Perez-Almendros, L. Espinosa-Anke, and S. Schockaert. Pre-training language models for identifying patronizing and condescending language: An analysis. *LREC*, 2022.

[128] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, 2019.

[129] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020.

[130] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, 2021.

[131] B. Plewes and R. Stuart. The pornography of poverty: A cautionary fundraising tale. *Ethics in action: The ethical challenges of international human rights nongovernmental organizations*, pages 23–37, 2007.

[132] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[133] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495, 2015.

[134] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30, 2016.

[135] C. Poth, J. Pfeiffer, A. Rücklé, and I. Gurevych. What to pre-train on? efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, 2021.

[136] I. Price, J. Gifford-Moore, J. Flemming, S. Musker, M. Roichman, G. Sylvain, N. Thain, L. Dixon, and J. Sorensen. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.15. URL https://www.aclweb.org/anthology/2020.alw-1.15.

[137] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[138] L. V. Ramos, A. M. Monterde, V. Pachón, and J. Mata. I2c at semeval-2022 task 4: Patronizing and condescending language detection using deep learning techniques. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 459–463, 2022.

[139] A. Ramponi and E. Leonardelli. Dh-fbk at semeval-2022 task 4: leveraging annotators' disagreement and multiple data views for patronizing language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 324–334, 2022.

[140] A. R. Rao. Asrtrans at semeval-2022 task 4: ensemble of tuned transformer-based models for pcl detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 344–351, 2022.

[141] C. Rastogi, N. Mofid, and F.-I. Hsiao. Can we achieve more with less? exploring data augmentation for toxic comment classification. *arXiv preprint arXiv:2007.00875*, 2020.

[142] R. Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

[143] S. Redfern and I. Norman. Quality of nursing care perceived by patients and their nurses: an application of the critical incident technique. part 2. *Journal of Clinical Nursing*, 8(4):414–421, 1999.

[144] P. Resnik. Four revolutions. *Language Log, February*, 5, 2011.

[145] R. Rezapour, S. H. Shah, and J. Diesner. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 35–45, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1305. URL `https://www.aclweb.org/anthology/W19-1305`.

[146] S. Rosenthal, N. Farra, and P. Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL `https://www.aclweb.org/anthology/S17-2088`.

[147] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.

[148] E. W. Said. *Orientalism*. Routledge & Kegan Paul Ltd., 1978.

[149] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[150] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence,

Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://www.aclweb.org/anthology/P19-1163.

[151] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://www.aclweb.org/anthology/2020.acl-main.486.

[152] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, 2022.

[153] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[154] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gambäck. SemEval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.semeval-1.99.

[155] D. Siegel. The failure of condescension. *Victorian Literature and Culture*, 33 (2):395–414, 2005.

[156] M. Siino, M. Cascia, and I. Tinnirello. Mcrock at semeval-2022 task 4: Patronizing and condescending language detection using multi-channel cnn, hybrid lstm, distilbert and xlnet. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 409–417, 2022.

[157] P. Simpson. *Language, ideology and point of view*. Routledge, 2003.

[158] A. Singh. Team lego at semeval-2022 task 4: Machine learning methods for pcl detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 369–373, 2022.

[159] S. Smiles. *Self-Help; with illustrations of character and conduct*. John Murray, 1866.

[160] G. J. Steen, A. G. Dorst, J. B. Herrmann, A. A. Kaal, and T. Krennmayr. Metaphor in usage. *Cognitive Linguistics*, 21(4), 2010.

[161] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[162] R. Straubhaar. The stark reality of the 'white saviour'complex and the need for critical consciousness: A document analysis of the early journals of a freirean educator. *Compare: A Journal of Comparative and International Education*, 45(3):381–400, 2015.

[163] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL `https://www.aclweb.org/anthology/P19-1159`.

[164] K. Tandon and N. Chatterjee. Team lrl_nc at semeval-2022 task 4: Binary and multi-label classification of pcl using fine-tuned transformer-based models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 421–431, 2022.

[165] Y. Tay, D. Ong, J. Fu, A. Chan, N. Chen, A. T. Luu, and C. Pal. Would you rather? a new benchmark for learning machine alignment with cultural values and social preferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5369–5373, 2020.

[166] O. Thomas-Olalde and A. Velho. Othering and its effects–exploring the concept. *Writing postcolonial histories of intercultural education*, 2:27–51, 2011.

[167] R. Van der Goot. Machamp at semeval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task multi-lingual learning for a pre-selected set of semantic datasets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1695–1703, 2022.

[168] T. A. Van Dijk. *The Role of the Press in the Reproduction of Racism*. Springer, 2012.

[169] T. A. Van Dijk. Critical discourse analysis. *The handbook of discourse analysis*, pages 466–485, 2015.

[170] C. Van Hee, E. Lefever, and V. Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, 2018.

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

[172] T. Vu, D. Q. Nguyen, and A. Nguyen. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341, 2021.

[173] A. Wang and O. Russakovsky. Directional bias amplification. In *International Conference on Machine Learning*, pages 10882–10893. PMLR, 2021.

[174] Y. Wang, Y. Wang, B. Ling, Z. Liao, S. Wang, and J. Xiao. PINGAN ominisinitic at SemEval-2022 task 4: Multi-prompt training for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 313–318, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.40. URL https://aclanthology.org/2022.semeval-1.40.

[175] Z. Wang and C. Potts. Talkdown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. URL https://www.aclweb.org/anthology/D19-1385.

[176] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[177] Z. Waseem, T. Davidson, N. Ithica, D. Warmsley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. *ACL 2017*, page 78, 2017.

[178] J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, 2019.

[179] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[180] C. C. Wilson and F. Gutierrez. Minorities and the media. *Beverly Hills, CA, London: Sage*, 1985.

[181] R. Wodak. *Language, Power and Ideology: Studies in political discourse*, volume 7. John Benjamins Publishing Company, 1989.

[182] R. Wodak. Critical discourse analysis. *Qualitative research practice*, 185: 185–204, 2004.

[183] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[184] J. Xu. Xu at semeval-2022 task 4: Pre-bert neural network methods vs post-bert roberta approach for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 479–484, 2022.

[185] W. Yin and A. Zubiaga. Hidden behind the obvious: misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210, 2022.

[186] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf`.

[187] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop*

on Semantic Evaluation, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010. URL https://www.aclweb.org/anthology/S19-2010.

[188] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.semeval-1.188.

[189] Y. Zhang, J. Baldridge, and L. He. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, 2019.

[190] D. Zhao, A. Wang, and O. Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14830–14840, October 2021.

[191] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

[192] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, 2019.

[193] J. Zhao, S. Mukherjee, S. Hosseini, K.-W. Chang, and A. Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.260. URL https://www.aclweb.org/anthology/2020.acl-main.260.

[194] X. Zhao and A. Rios. Utsa nlp at semeval-2022 task 4: An exploration of simple ensembles of transformers, convolutional, and recurrent neural networks. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 379–386, 2022.

[195] N. Zhou and D. Jurgens. Condolences and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, 2020.

[196] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76, 2020.

[197] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences, 2020.

[198] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.