UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA Y SISTEMAS DE DATOS

TRABAJO FIN DE GRADO

DEVELOPMENT OF A MORAL FOUNDATIONS ESTIMATION SYSTEM BASED ON NATURAL LANGUAGE PROCESSING TECHNIQUES AND TRANSFORMER MODELS

ANNY D. ÁLVAREZ NOGALES

JUNIO 2024

TRABAJO DE FIN DE GRADO

Título:	Desarrollo de un Sistema de Estimación de Fundamentos Morales basado en Técnicas de Procesamiento de Lenguaje Natural y Modelos Transformers
Título (inglés):	Development of a Moral Foundations Estimation System based on Natural Language Processing Techniques and Transformer Models
Autor:	Anny D. Álvarez Nogales
Tutor:	Óscar Araque Iborra
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

DEVELOPMENT OF A MORAL FOUNDATIONS ESTIMATION SYSTEM BASED ON NATURAL LANGUAGE PROCESSING TECHNIQUES AND TRANSFORMER MODELS

Anny D. Álvarez Nogales

Junio 2024

Resumen

Los valores morales son principios fundamentales y convicciones que guían cómo las personas actúan y se relacionan con los demás, influyendo en sus decisiones y comportamientos éticos diarios. Comprender las limitaciones en la detección de estos valores es crucial, especialmente en medios digitales, donde la interpretación de la moralidad puede ser más compleja debido a la diversidad de contenido y contexto. Las plataformas digitales han revolucionado la forma en que las personas se comunican e interactúan, generando una creciente necesidad de asegurar que el contenido compartido sea adecuado. Identificar estos valores puede permitir entender mejor las intenciones y mensajes subyacentes en el contenido, haciendo más conscientes a las personas de cómo este puede influir en la percepción y toma de decisiones.

Este trabajo se centra en la evaluación del desempeño de modelos basados en transformadores BERT y RoBERTa, que representan el estado del arte en el procesamiento del lenguaje natural (PLN) en una variedad de aplicaciones. En particular, se investiga la capacidad de estos modelos para detectar la moralidad en textos utilizando los fundamentos éticos definidos en la Teoría de los Fundamentos Morales (MFT), que distingue cinco rasgos morales y diferencia entre vicio y virtud. Además, se analizan diferentes niveles de complejidad en la detección de la moralidad y se explora el impacto en los modelos al incorporar información subjetiva y detalles adicionales mediante el uso de recursos léxicos que reflejan emociones y moralidad. Finalmente, se examina cómo estos enfoques se desempeñan en diferentes dominios y de qué manera benefician la comprensión del texto.

Los resultados obtenidos resaltan que la inclusión de estos léxicos, aunque depende de la complejidad de la tarea, tiene un impacto positivo en la capacidad de los modelos para distinguir la moralidad subyacente en el texto. Esta mejora se observa tanto en situaciones similares al dominio de entrenamiento como en dominios distintos, lo que evidencia la efectividad de los enfoques que integran datos enriquecidos con perspectivas subjetivas para aumentar la robustez de los modelos.

Palabras clave: Valores Morales, Procesamiento de Lenguaje Natural, Transformers, Modelos de Lenguaje, Bert, Roberta

Abstract

Moral values are fundamental principles and convictions that guide how people act and relate to others, influencing their daily ethical decisions and behaviours. Understanding the limitations in the detection of these values is crucial, especially in digital media, where the interpretation of morality can be more complex due to the diversity of content and context. Digital platforms have transformed the way people communicate and interact, creating a greater need to ensure that the content shared is appropriate. Identifying these values can allow for a better understanding of the intentions and underlying messages in content, making people more aware of how it can influence perception and decision-making.

This work focuses on the evaluation of the performance of models based on BERT and RoBERTa transformers, which represent the state of the art in natural language processing (NLP) in a variety of applications.In particular, the ability of these models to detect morality in texts is investigated using the ethical foundations defined in Moral Foundations Theory (MFT), which identifies five moral traits and distinguishes between vice and virtue. It also analyses different levels of complexity in morality detection and explores the impact on the models of incorporating subjective information and additional detail through the use of lexical resources reflecting emotion and morality. Finally, it examines how these approaches perform in different domains and how they benefit text comprehension.

Results show that the addition of these lexicons, despite depending on the complexity of the task, positively influences the models' ability to distinguish the underlying morality in the text. Results improvement is observed in both situations, similar to the training domain and in different domains, demonstrating the effectiveness of approaches that integrate enriched data with subjective perspectives to increase the models' robustness.

Keywords: Moral values, Natural Language Processing, Transformers, Language Model, Bert Model, Roberta Model

Agradecimientos

Deseo expresar mi más sincero agradecimiento a todas las personas que, de una manera u otra, han contribuido al desarrollo de este trabajo durante los últimos meses.

En primer lugar, quiero agradecer profundamente a Óscar por aceptar la responsabilidad de ser mi tutor, brindando su confianza durante este proceso de investigación. Su dedicación y apoyo han sido esenciales para llevar a cabo este trabajo.

Gracias también al Grupo de Sistemas Inteligentes, por recibirme con los brazos abiertos y por su constante apoyo. Su experiencia y ánimo han sido de gran valor; sois personas maravillosas y me he sentido muy acogida, como en una familia.

No puedo olvidar el enorme respaldo brindado por mis amigos. Sofía, gracias por hacer estos días en la escuela más llevaderos, me llevo a una hermana de aquí. Y por supuesto, quiero expresar mi más profundo agradecimiento a mi madre y a mi hermana por su apoyo incondicional, espero que estéis orgullosas de mí.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

Digital platforms have transformed the way we communicate and share information, becoming a habitual part of our daily lives. With the vast amount of linguistic and multimedia data published daily on the Internet, we are constantly exposed to a flow of information that can influence various aspects of our lives, from our online behaviour and interaction with other users to our perception of reality and participation in the public sphere [65].

Platforms such as Twitter, Reddit and portals news act as daily channels for people to share their views, ideas and to stay informed about global issues [30]. However, this access to information could lead to debates and conflicts that could be very divisive and contain biased justifications [74]. The presence of sensitive topics such as vaccination [20], climate change or political debates tends to provoke divergent opinions and elicit strong reactions from people influenced by their environment and political preferences [20, 94]. Indeed, online conversations that may reflect different values, ideologies or political affiliations also take place on these platforms [44, 17, 93, 47], which can lead to phenomena such as radicalisation or religious and political hate [92].

It's fundamental to recognize that these matters are related to individual moral prin-

ciples that differ from person to person, influencing in how we perceive and address these issues [80]. These differences, shaped by culture and historical context, generate a variety of viewpoints that contribute to polarisation in public debate [16]. There is where the importance of moral values awareness come into action. These are defined as innate, emotional foundations that guide our perceptions, ethical decisions and influence social judgements [33]. Social psychologists, such as Jonathan Haidt and Joseph Graham, claim that humans have five innate moral foundations that influence our choices [35], leading to the definition of Moral Foundations Theory [32]. The MFT proposes, initially, five moral foundations. Each foundation expresses a vice and a virtue polarity: (1) the care/harm foundation which deals with the sensitivity towards the suffering of others; (2) fairness/cheating covering aspects of reciprocity and motivations to be fair; (3) loyalty/betrayal covering aspects of in group cooperation, and the intuition of being loyal to one's group; (4) authority/subversion is related to the innate intuition of endorsing hierarchies that we find just; (5) the purity/ degradation foundation which deals with our innate drive of preferring cleanliness of body and soul over hedonism. Recently, the foundation (6) liberty/oppression was added, which deals with the feelings of reaction and resentment that people feel towards those who dominate them and restrict their freedom.

Since people use these platforms to express their beliefs and opinions, social media are loaded with very different moral content, values and beliefs, which can generate arguments and a hostile tone [22] when any of the sensitive issues, mentioned above, come into play.

Due to the language used in these discussions can reveal information about the moral values of the individual who wrote it [46, 31], ability to detect through the use of intelligent systems makes it possible not only to study phenomena such as opinion formation and improve online communication, but also to create more ethical and responsible intelligent systems, a central concern today [55, 87, 83, 72].

However, such morality detection can face challenges due to the complexity of human language, the lack of domain knowledge and annotated resources in different contexts. Classifiers are often trained without taking into account that moral values are related and vary according to the domain or domain of discussion [60, 59]. These limitations may interfere with the ability of language models to fully capture the diversity and complexity of moral expressions in different situations and domains, making it difficult to understand these values.

1.2 Research Questions

This section outlines the fundamental research questions guiding the project, focusing on the influence of task complexity on moral prediction in text and the integration of subjective information into learning models. It also outlines the overall aims of the project in addressing these questions and advancing the understanding of automated moral estimation. The paper aims to answer the following research questions:

RQ1: How does varying levels of task complexity impact the accuracy and performance differences between two models used for moral prediction in text?; this question delves into the influence of task complexity on the effectiveness of models in predicting moral values in text. By assessing performance across different complexity levels, the study seeks to understand how training efficiency is affected and identify strategies to mitigate negative impacts. Following this, it is also study **RQ2: What is the effect of integrating subjective information, such as emotions and moral knowledge in text, on the generalization capacity of learning models used for automated moral assessment?**; that explores the impact of incorporating information into learning models on their ability to generalize moral evaluation. By quantifying the generalization capacity, the study aims to evaluate the effectiveness of the model's performance.

In addition to these primary research questions, related with moral estimation, the project also incorporates the following (detailed in [71]): **RQ3: To what extent can the diversity of views in moral annotations be useful for automated moral assessments?**; that investigates the usefulness of incorporating diverse perspectives from moral annotations in improving automated moral estimators and **RQ4: Is it possible to automatically assess whether a text is challenging to annotate?**; to investigate the feasibility of automatically determining the difficulty of annotating a text based on the level of disagreement between annotators.

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

- ***Chapter 1*** provides an overview of the context of the problem, the main aims of the research and the structure of this work.
- ***Chapter 2*** presents the technologies and libraries resources that enabled to carry on

the implementations.

- ***Chapter 3*** details the state of the art in technologies and techniques previously used for moral detection in text.
- ***Chapter 4*** presents in detail the research proposals suggested addressing the identified problem and the defined objectives.
- ***Chapter 5*** describes the resources, including models and the metrics used, as well as a description of the implementation of the experiments.
- ***Chapter 6*** offers a detailed description of the results obtained.
- ***Chapter 7*** presents general conclusions and future work.
- ***Chapter 8*** and ***Chapter 9*** provide details about the ethical, economic, social, and financial aspects.

CHAPTER 2

Enabling Technologies

2.1 Libraries and Frameworks

This project was implemented using the Python programming language [11]. It is an interpreted, object-oriented, high-level programming language with dynamic semantics and is the most widely used computer programming language, especially in the fields of data science and machine learning. Its extensive ecosystem includes various libraries for processing, manipulating and visualising all types of data. Some notable libraries include NumPy for numerical computing, NLTK for natural language processing, Pandas for data manipulation and analysis, and Seaborn for statistical data visualisation.

To carry out the different experiments, the following libraries have been used:

Pandas [8]: library that simplifies the tasks of data manipulation in Python. It provides easy to use data structures and tools for loading, cleaning, transforming and preparing structured datasets for modelling. Pandas is built on top of two core Python libraries: matplotlib for data visualization and NumPy for mathematical operations; it also acts as a wrapper over these libraries, allowing the access many of matplotlib's and NumPy's methods with less code. Some key features offered by the project are:

- Creation and manipulation of dataframes for Tabular and CSV data
- Tools for handling missing data, duplications, formatting issues etc.
- Apply mathematical operations for fast data transformation

MatplotLib [7]: is a widely-used Python library for data visualization. It provides a wide range of functions for creating static plots, interactive plots, scatter plots, histograms, bar charts, among other types of visualizations. Matplotlib is flexible and powerful, making it a fundamental tool for data exploration and communication in fields such as data science, engineering, academic research, and many others. Moreover, his primary purpose is to provide the tools to represent data graphically, making it easier to analyze and understand.

Seaborn [10]: it is also a Python library used to plot graphs using Matplotlib, Pandas and Numpy. It is built on top of Matplotlib and helps to visualise univariate and bivariate data, making it easier to create. It uses a variety of themes to decorate Matplotlib graphs and is commonly used for data science and machine learning tasks, providing a high-level interface for drawing attractive and informative statistical graphs. Additionally, Seaborn offers support for complex visualizations such as multi-plot grids and categorical plots, making it a versatile tool for data analysis and exploration

HuggingFace Transformers [6]: is an open source Python library that provides access to thousands of pre-trained Transformers models for natural language processing (NLP), computer vision, audio tasks and more helping users to build, deploy and train machine learning models. Founded in 2016 by French entrepreneurs Clément Delangue, Julien Chaumond, and Thomas Wolf, the company originally developed a chatbot app of the same name for teenagers, but is evolving into a hosting platform for natural language processing and machine learning domains; it simplifies the process of implementing Transformers models by abstracting the complexity of training or deploying models in lower-level ML frameworks such as PyTorch, TensorFlow, and JAX. This resource hosts a large collection of open source machine learning models and datasets, and provides access to the models implemented in this project.

Scikit-learn (Sklearn) [9]: is an open-source machine learning and data modeling library for Python. It provides a selection of efficient tools for machine learning and statistical modeling, including classification, regression, clustering, and dimensionality reduction. This library is built upon NumPy, SciPy, and Matplotlib, allowing for data manipulation and preprocessing for experiments.

Ekphrasis [15]: is a text processing tool designed specifically for text from social net-

works, such as Twitter or Facebook. Ekphrasis performs tokenization, word normalization, word segmentation (for splitting hashtags), and spell correction, using word statistics from two large corpora (English Wikipedia and 330 million English tweets from Twitter). This tool allows the preprocessing and adaptation of raw data into a more suitable form for machine learning models.

MoralStrength [12]: framework that contains functions to access to MoralStrength lexicon, which enables researchers to extract the moral valence from a variety of lemmas in texts.

Autorank [36]: is a simple Python package with functions to simplify the comparison between (multiple) paired populations and to automatically compare populations defined in a block design data frame. There are five variants of statistical tests a Bayesian signed-rank test, a paired t-test, a Wilcoxon or repeated measures ANOVA with Tukey's HSD as a post-hoc test, depending on the data distribution and the number of classifiers.

CodeCarbon [5]: is an open source tool designed to track and calculate the carbon impact associated with computational workloads.

The project environment used was **Jupyter Notebook**, which is a web interpreter for creating and sharing interactive documents containing executable code, equations, visualisations and narrative text. It is especially useful for experimentation and presentation of analysis, allowing separate code snippets to be run and results to be viewed immediately. The Jupyter environment is used through **Jupyter Hub**, provides an interface that supports multiple documents, such as notebooks, and tools in a single window, allowing the management of complex projects. An overview of the structure is shown in Figure 2.1.

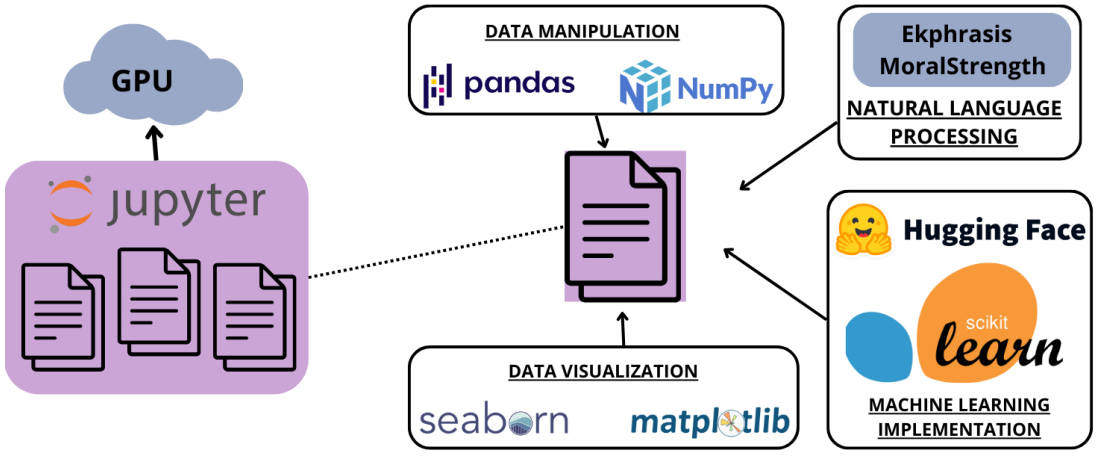

Figure 2.1: Project technologies overview

2.2 Transformer Models

Prior to the arrival of Transformers in 2017, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were the predominant methods in deep learning.

CNNs are structured with layers of convolutional filters that learn hierarchical representations of features in images or text. These filters are applied across input data to extract local patterns, such as word embeddings or n-gram features, which are then aggregated to form higher-level representations. However, CNNs are primarily used for image-related tasks.

RNNs are characterised by connections that form directed cycles, enabling them to retain memory of past inputs. This recurrent structure allows RNNs to capture temporal dependencies in sequential data, making them suitable for tasks such as speech recognition, language modelling and machine translation. However, they face challenges in understanding long-range dependencies and maintaining memory over long sequences of text.

New architectures like Long Short-Term Memory (**LSTM**), address some of these limitations by introducing a memory cell mechanism. This mechanism enhances their ability to capture and retain information over longer sequences more effectively. However, even with greater memory capacity, LSTM networks still encounter challenges in effectively handling large volumes of text data (Figure 2.2).

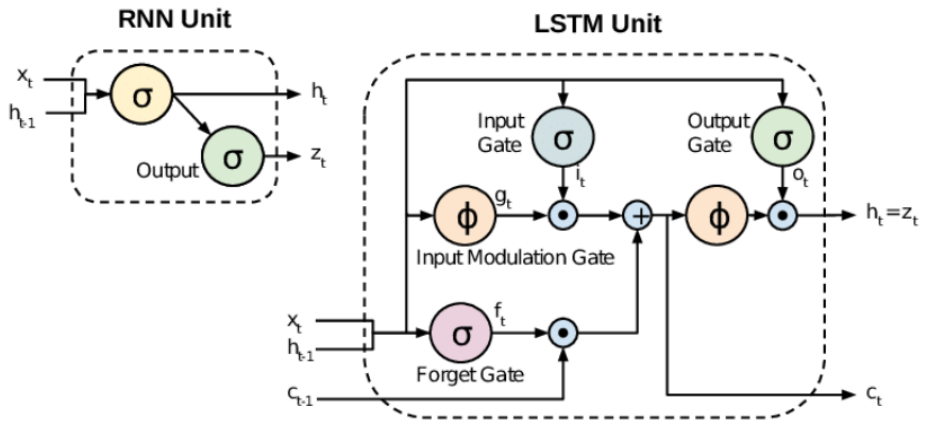

Figure 2.2: RNN vs LSTM architecture [1]

In 2017, a breakthrough occurred with the introduction of a family of models called **'Transformers'** which belong to the category of **Large Language Models (LLM)**. These models transform natural language processing by relying entirely on the 'attention'

mechanism for processing large amounts of sequential data [90].

2.2.1 Architecture

Transformer models revolutionized the field of Natural Language Processing (NLP) and machine learning by introducing novel mechanisms that significantly enhance the handling of sequential data. Figure 2.3 illustrates the overall architecture of these models.

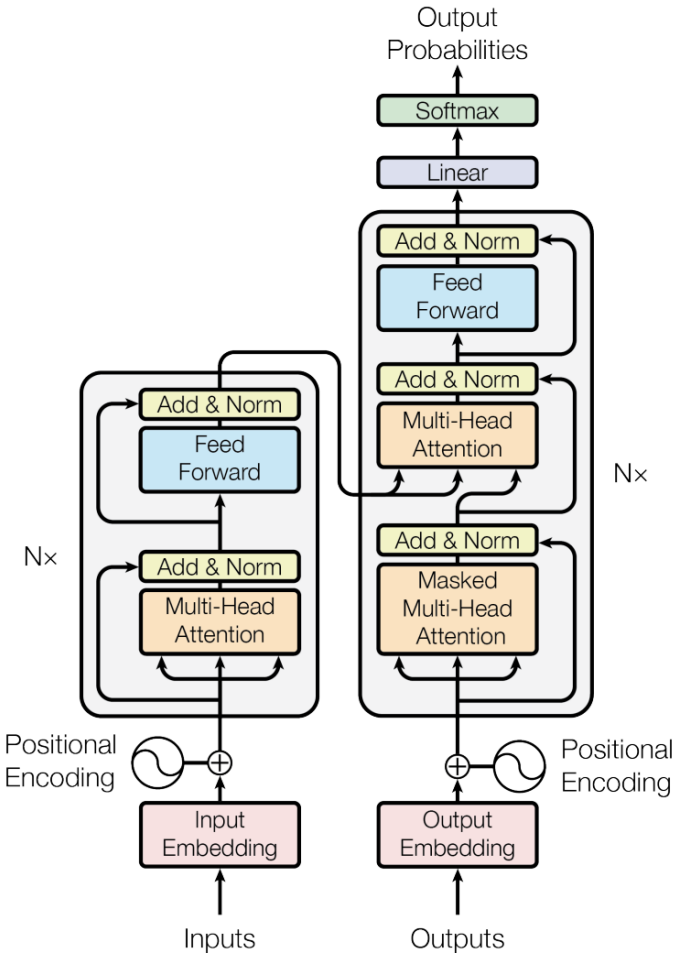

Figure 2.3: The Transformer Model architecture [90]

The key components of Transformer models are:

- **Embeddings**: these are vectors that represent input tokens or words, capturing their semantic meanings and allowing the model to understand the contextual information of each token within the input sequence.
- **Positional encodings**: positional encodings are added to the input embeddings to

provide information about the position of each token within the input sequence. Since Transformers do not inherently understand the sequential order of tokens, positional encodings help the model distinguish between tokens based on their position in the sequence.

- **Self-Attention Mechanism**: 'self-attention', also known as 'auto-attention', is a mechanism that allows the model to weigh the importance of different words in the input sequence based on their relevance to each other.

- **Multi-Head Attention**: 'multi-head attention' extends the 'self-attention' mechanism by allowing the model to attend to different parts of the input sequence simultaneously. This technique that allows the model to pay attention to different parts of the input sequence simultaneously, and capture patterns of long-range dependencies in the input sequence (Figure 2.4).

- **Scaled Dot-Product Attention**: it is used to calculate the importance of each word relative to other words in the sequence.

- **Feed-Fordward Network**: a dense layer of neural networks that is applied after the 'multi-head attention' mechanism in each block of the Transformer network, found in both the encoder and decoder layers. This layer consists of two linear transformations, each followed by a non-linear activation function, such as the ReLU activation function. Applied independently to each position in the sequence, this network enables the model to capture complex patterns and relationships within the data, transforming the vectors into a representation more suitable for the output task.

- **Encoder**: it consists of multiple layers, each containing a 'self-attention' mechanism followed by a feed-forward network. The 'self-attention mechanism' enables the encoder to capture dependencies between tokens in the input sequence, while the feed-forward network helps to refine these representations further. The output of the encoder is a sequence of 'context-aware' embeddings for each token in the input.

- **Decoder**: it is similar in structure to the encoder but includes an additional attention mechanism called 'encoder-decoder attention'. This attention mechanism allows the decoder to attend to relevant parts of the input sequence (encoded by the encoder) while generating the output sequence. The output of the decoder is a sequence of tokens representing the generated output sequence.

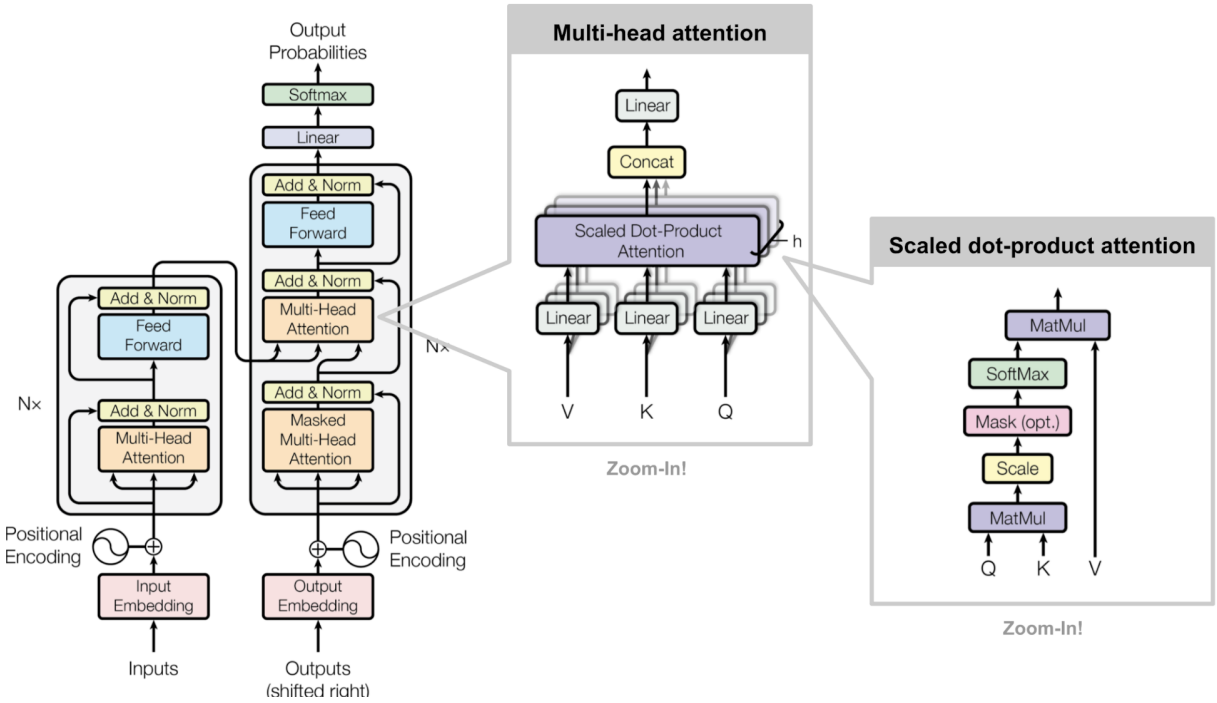

Figure 2.4: Multi-Head Attention component [1]

2.2.2 BERT and RoBERTa Models

Bidirectional Encoder Representations from Transformers **(BERT)** is a **pre-trained language model** developed by Google in 2018 [25]. It was trained with large text corpora (3.3 billion words) unlabelled to learn to predict missing words in a given text sequence. During the training process, it learns to capture the contextual information of words in a sequence, allowing it to understand words in context.Unlike the original Transformer architecture which is trained in only one direction (where each token/word only attended to the previous tokens in the self-attenuation layers), BERT is trained in a bidirectional manner, meaning that it can capture the context of words in both directions. Other differences are:

- **Masking**: BERT uses masked language pre-training (Figure 2.5), to obtain a more detailed word representation. Some words in a training text are masked, and it predicts the masked word based on the surrounding context.

- **Multi-task classification**: it can perform a variety of NLP tasks such as text classification, inference or question answering, using fine-tuning (Figure 2.6).

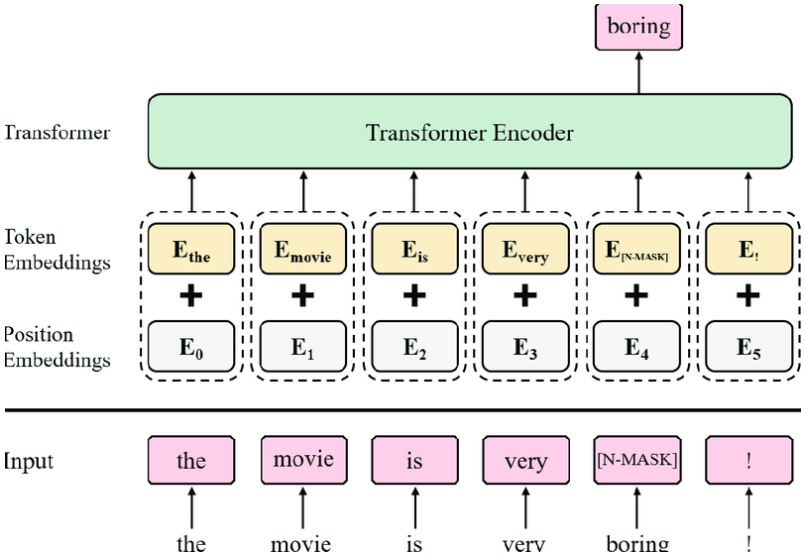

Figure 2.5: BERT Masked Language Modelling [90]

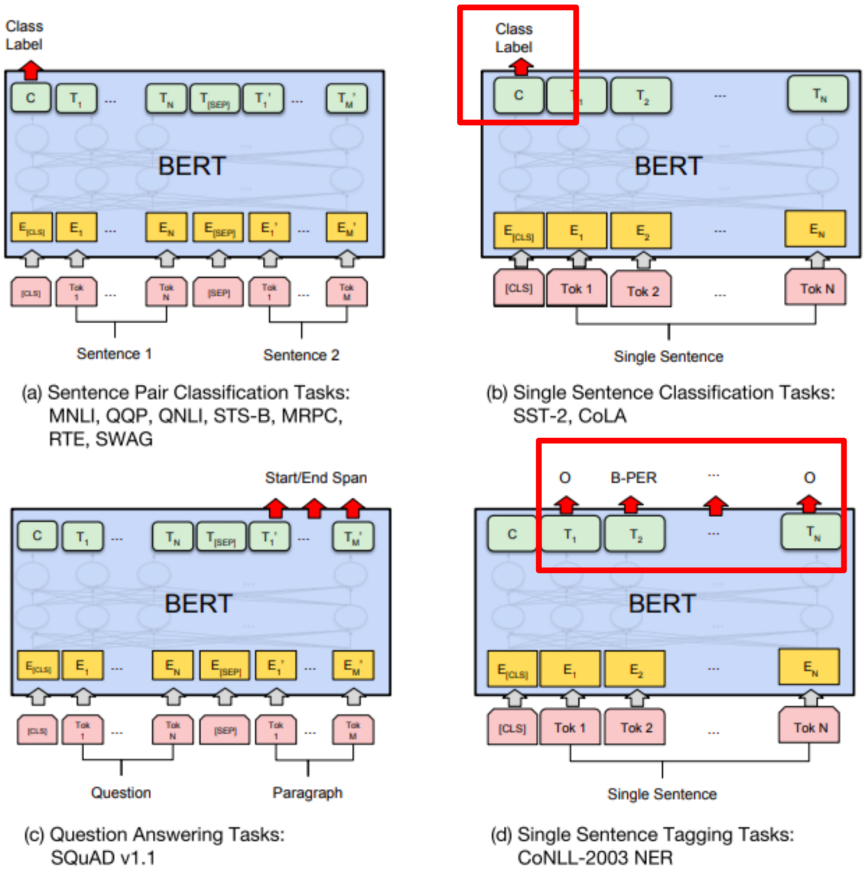

Figure 2.6: BERT multitask architecture [90]

Figure 2.7 shows the architecture of the BERT model, which consists of **only one bi-directional** multi-layer transformer **encoder**. There are two versions of it: BERT Base, the smaller and faster version with 12 layers of multi-headed care and 110 million parameters and BERT Large, a more powerful version involving more computational and storage costs with 24 layers of multi-headed and 340 million parameters.

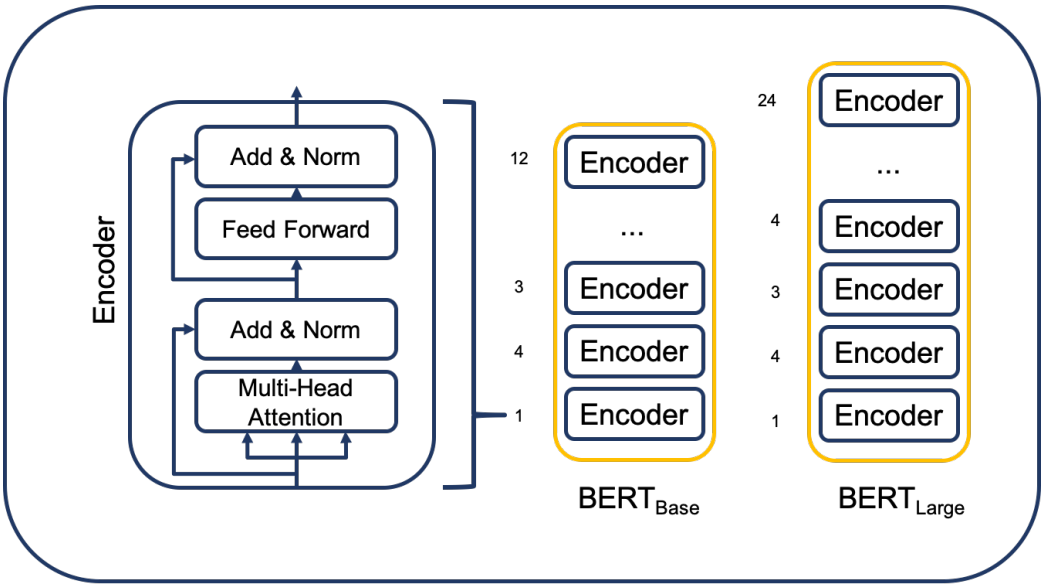

Figure 2.7: BERT architectures, based on the original implementation in [90]

A Robustly Optimized BERT Pretraining Approach **(RoBERTa)**, is a pre-trained language model developed by Facebook AI Research in 2019 [63]. The research team realized that BERT was not trained with sufficient amount of data. Thus, the main new feature of RoBERTa is that it uses a larger training dataset, specifically 160 GB of text instead of the 16 GB that was originally used to train BERT. Other improvements were also added such as longer training (increasing the number of iterations) and on new datasets (including text extracted from the Internet), the elimination of the next sentence prediction task or the change of the masking pattern to a dynamic one, whereby different sets of tokens are masked at each training iteration. This helps prevent overfitting to specific patterns in the training data and improves the model's ability to generalize to new data. Table 2.1 shows a summary of the characteristics of both models:

	BERT	**RoBERTa**
Developed by	Google (2018)	Meta AI (2019)
Base Architecture	Bidirectional Transformer (encoder only)	Bidirectional Transformer (based on BERT)
Tokenizer	WordPiece tokenizer	SentencePiece tokenizer
Training Corpus	English Wikipedia BooksCorpus	English Wikipedia BooksCorpus CC-News, OpenWebText
Parameters	BERT base:110M BERT large: 340M	RoBERTa base: 125M RoBERTa large: 355M
Word Masking	15% of words are randomly masked during training	More random and dynamic masking during training

Table 2.1: Comparison of Pre-trained Language Models: BERT vs RoBERTa.

CHAPTER 3

Related Work

This section provides an overview of the different approaches used to detect morality in textual data. It describes traditional methods as well as the most recent approaches based on transformer models. Traditional approaches include Natural Language Processing (NLP) and specific lexicon based methods, while the most advanced methods exploit the contextual capabilities of transformer based models, such as embeddings. In addition, this section will cover the methods used to acquire domain knowledge to improve moral prediction in text.

3.1 Feature Extraction Methods

Feature extraction methods are fundamental in the field of Natural Language Processing and Artificial Intelligence, as they are responsible for representing data for models for tasks such as the classification of moral values in text. These approaches are designed to identify and highlight specific aspects of textual information that are relevant to the task. These methods can generally be categorized into 'traditional feature extraction techniques' and 'modern embedding models' to transform raw textual data into a numeric form understandable by machines. This work is framed within the Moral Foundations Theory (MFT), which emphasizes that despite the idea that morality varies significantly between individuals due

to differences in historical, cultural, and political contexts, there is 'an intuitive' basis suggesting that humans possess an innate sense of ethics [34]. This theory has been widely accepted in computational linguistics works and identifies five main traits of morality:

- **Harm/Care**: which deals with the sensitivity towards the suffering of others. This moral trait involvers people's concern for the well-being of others. This intuition involves our natural impulses to protect our own children and others from harm. It includes virtues such as compassion, kindness and caring for the vulnerable.
- **Fairness/Justice**: covering aspects of reciprocity and motivations to be fair. This dimension relates to dealing people fairly, maintaining justice and upholding the rights of citizens. It may also include the protection of personal autonomy and the idea that people should be rewarded in proportion to what they contribute.
- **Loyalty/Betrayal**: covering aspects of in-group cooperation, and the intuition of being loyal to one's group. This involves a person disposition to show loyalty and commitment to an individual, group or cause. This moral trait implies an emotional connection and attachment to the group to which one belongs, as well as a distrust of those outside the group. Loyalty is manifested through actions that benefit the group and may involve personal sacrifice, while betrayal is manifested by acting to the detriment of the group or its interests.
- **Authority/Subversion**: which is related to the innate intuition of endorsing hierarchies that we find just. It includes instinctive respect for hierarchy, obedience to legitimate authority, duty, awe and admiration for those in power, as well as veneration of traditions. This foundation incorporates the requirement that respect must be shown to parents, teachers and others in positions of authority.
- **Purity/Degradation**: which refers to the valuing of cleanliness and purity over pollution and degradation, both physically and morally. This trait is formed from the psychology of disgust and pollution, and involves striving to live in an elevated and noble manner, avoiding immoral activities that may pollute the body and soul. It relates to self-discipline, spirituality and respect for the body as a temple.

Recent developments encouraged the need to add a moral foundation to reflect the presence or absence of freedom in text, adding a new moral feature to the MFT [41].

- **Liberty/Oppression**: which deals with the feelings of reaction and resentment that people feel towards those who dominate them and restrict their freedom.

Figure 3.1: Moral Foundations Theory [2].

Previous work on detecting moral values in texts was developed using **statistical approaches** for text feature extraction, such as word count based on dictionaries like the Moral Foundations Dictionary (MFD) [27] or its extension, the Extended Moral Foundations Dictionary (eMFD), where different moral dimensions were analysed using word count techniques [77]. These studies analysed various moral dimensions using word count techniques, focusing on analysing the presence and frequency of keywords associated with different moral dimensions in a text. The dictionaries used contain lists of words or lemmas previously identified as indicators of moral traits described in the MFT, and counts are simply used as features in supervised classifiers.

Other works developed **unsupervised approaches** trying to overcome the issues related to the simple counts of lexicon words, embed the moral values in continuous spaces using word embeddings. Word embeddings [45] represent each word as a real-valued vector intended to represent his meaning, words that are close in this vector space are expected to be similar in meaning and these can be learned by training a model to complete text fragments [67, 68]. Using this approach, [69] classifies the morality of online news headlines without requiring previously annotated data, using embeddings and a low-dimensional feature representation to characterise the text with respect to a set of target words. Text is scored along moral dimensions, calculating bias and framing intensity scores for each moral foundation using the method of 'Frame Axis', that consist on projecting words onto micro-frame dimensions characterised by opposing sets of words, representing the text with moral dimension scores, and allowing the automatic identification of relevant keywords or concepts in a text (see more in [53]).

The use of **word embedding models**, has addressed the challenges of capturing context, bringing out the semantic similarity of words that captures different facets of the

meaning of a word, aspects that statistical techniques struggled to effectively represent. Models vary in their methodologies, some of them like Word2Vec embeddings [67] are derived from training a shallow feedforward neural network, whereas others like GloVe [75], embeddings are acquired through matrix factorization techniques. Using this approach, embeddings models are often pre-trained on large amounts of data and then fine-tuned to task-specific sets in a process called Transfer Learning' [88]. These embeddings are then used in **training supervised classification models**, as demonstrated in [78], which investigates the relationship between basic principles of human morality and the expression of opinions in user-generated text using two methods: a lexicon or dictionary approach based on counts (described below) and deep learning classification using embeddings and supervised learning models like Support Vector Machine (SVM) [26], Random Forest (RF) [64], and Long short-term memory (LSTM) [37]. Using a similar approach with word embeddings extracted from text, [49] use them as input in an ElasticNet regression model to predict each MF score and investigate the relationship between language use and moral concern in status social media posts. Moreover, [12] uses embeddings extracted from different methods, such as statistical-based and cosine similarity, to feed regression models and predict the moral value reflected in tweets.

3.2 Transformer models for morality detection

Approaches based on word embeddings have been shown to be highly effective in various natural language processing tasks by **capturing the semantics of text**. However, these approaches often lack context, which can limit their ability to understand morality in text. This is where context embeddings come in; these are obtained from pre-trained language models, such as large language models, and address this limitation by capturing the broader linguistic context. In fact, recent embeddings advancements in Large Language Models (LLM) have shown that this architecture is particularly suitable for tasks related with morality [96, 91]. **LLMs** are machine learning models built upon deep neural networks with a large number of parameters, millions or even billions. These models have been pre-trained on large corpora of unlabelled text and can then be fine-tuned, to datasets to perform concrete natural language processing tasks, such as machine translation, text generation, automatic summarisation, sentiment analysis or text classification. In fact, these models have been shown to be able to incorporate even human prejudices about what is considered right or wrong, capturing general features of morality [81].

Transformer models are an example on the large list of LLM models. The introduc-

tion of these models in NLP tasks, in particular the emergence of the Bidirectional Encoder Representations from Transformers (BERT) [25] model, resulted in even better performing contextual embeddings.

The family of architectures **BERT** and its variants have demonstrated a more precise moral compass compared to previous embedding methods [82], enabling their use in tasks such as morality classification in user-generated texts from online sources like Facebook [49]; or even in specific contexts like politics [79], where tweets on the different positions towards US politicians on the issues of 'Gun Control' and 'Immigration' were studied, finding notable differences in the use of moral grounds by different political parties. It has also been employed in moral and classification tasks, considering various perspectives [54], or to understand how these values may vary across different domains on online platforms such as Reddit and Twitter, modelling different approaches to detecting the presence of morality [18, 76]. Finally, other works have gone further to analysing different contexts and develop new methods, such as a neural adaptation framework that uses instance weighting, to improve classification tasks across different domains [40].

3.3 Additional Knowledge into Large Language Models

However, despite the better performance of transformer-based models compared to traditional learning approaches, they face generalization problems due to being trained on specific data. There are significant variations in how moral values are expressed and discussed across different contexts, directly impacting the classifiers' ability to identify and categorize these values [57, 58]. In general, due to high subjectivity, morality prediction models do not achieve very high accuracy, but this does not mean that the model is incapable of correctly scoring the ratings, it simply means that the model interprets the ratings and associates a score to them as any person would. **Each model is linked to the context in which it has been trained**, a model that has been trained using a specific dataset will get consistent results when analysing related texts, whereas if the text is unrelated to the training set, the results will be more inaccurate. Techniques such as integrating supplementary information, like lexicons, have proven to enhance the accuracy of models in tasks such as sentiment classification [52] or Named Entity Recognition [98]. Other works more similar to this, integrate features that reflect the original text in the input of the model before processing, such as moral dictionaries (MFD) [78, 56] or lexicons like MoralStrength [76], for improved prediction accuracy in morality detection.

One of the most subtle strategies for incorporating this additional information is through

prompts methods. **Prompts** are essentially templates used alongside the original text to guide the model in performing a specific task, such as answering questions, completing sentences or text classification. In addition to guiding the model, prompts can incorporate **domain-specific knowledge** to enhance the interpretation of the data. For example, in text classification tasks in specialised areas such as hate speech, they can include key terms or contextualised phrases that help the model to better understand and classify the content.

In recent years, prompted learning has been used to mitigate objective differences between pre-trained language models and subsequent tasks, as well as to maximise the transferability of language models. One of the biggest successes has been in few-shot classification learning, where classifiers learn with only a few labelled examples of each class. The most challenging task is the generation of the appropriate prompt, as there is no universal template for all NLP tasks. Some hand-crafted prompts have been explored in tasks such as sentiment classification [62], where Chen and Zhang proposed a question-based approach that links label-related questions to each candidate sentence to help language models.

Given the success of related projects in prompting, such as the one by Jiang et al. [42] which introduced the first prompt-based sentence embedding method, or the study of [43] which showed that base models can achieve predictions comparable to large-scale language models using the prompt approach; this work has adapted the prompting method.

CHAPTER 4

Proposals

This section describes the methodology and key concepts necessary to understand the project. To answer the questions described in 1.2, the first step involves investigating how complexity affects the modeling of the morality detection task in two transformer models, considering preprocessing the data before inputting it into the model. In addition, the impact on model accuracy predictions of incorporating additional subjective knowledge reflected in the text, thereby increasing the complexity of the input, will be evaluated. This work will result in a moral estimation system as reflects in Figure 4.1

4.1 Complexity of morality detection

First approach focuses on the use of underlying morals reflected in text following the Moral Foundations Theory (MFT), to model the diversity of moral interpretations. Four classification tasks are defined, reflecting different levels of complexity in morality detection. These are designed to capture both, the presence and polarity of moral foundations, differentiating between moral virtues and moral vices. Table 4.1 summarises the tasks performed.

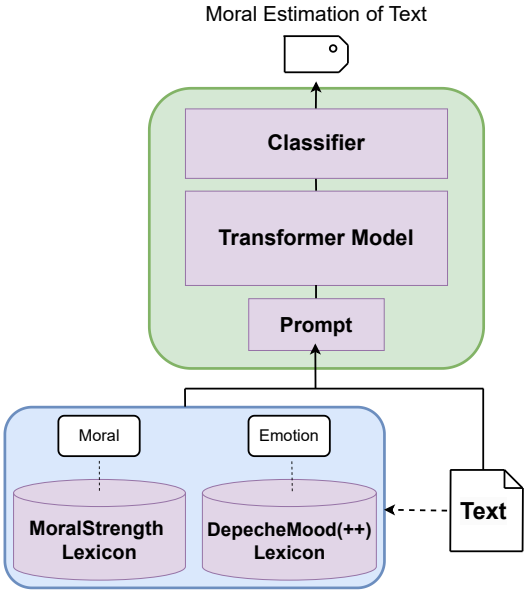

Figure 4.1: Final system overview.

Task	Definition
MPres	**Moral Presence**: detect the presence of each of the moral trait, determining whether each of the moral features of the MFT is present, considering one at a time.
MPol	**Moral Polarity**: infer the polarity or extreme present in moral traits. In this task, an additional level is added to the prediction, not only by distinguishing if a specific moral foundation is present,but also by identifying what polarity it has, distinguishing between vice and virtue.
MultiPres	**Multi Moral Presence**: focuses on identifying which of the moral features is most prominent in the text. All moral traits are considered, but only the predominant moral trait in the text is inferred.
MultiPol	**Multi Moral Polarity**: a further level of complexity is introduced by specifying whether this trait is presented as a virtue or a vice, thus providing a more detailed assessment of the morality of the text.

Table 4.1: Definition of the different tasks of text morality detection.

4.2 Knowledge Addition

Addressing the different levels of moral detection, it is explored how the inclusion of additional subjective information can vary the predictive capacity of the model in terms of morality. Specifically, data related to morality and emotions present in texts will be incorporated by making use of linguistic resources such as the lexicons: MoralStrength [13] and DepecheMood(++) [14].

- **MoralStrength**: is an extension of the Moral Foundations Dictionary (MFD), specifically designed to capture the moral rhetoric **based on the five moral traits** defined in MFT. This lexicon was manually enriched with synonyms extracted from WordNet, providing a list of lemmas that offer detailed information on the moral valence of each term, allowing for a more complete understanding of the moral dimension present in the text.
- **DepecheMood**(++): this extended version of the DepecheMood lexicon [85] provides a wide variety of lemmas, each associated with detailed information about **six different emotions**. For each emotional dimension, a score is provided indicating the relative likelihood of that emotion in the term, allowing the emotional nuance present in the text to be captured more accurately.

In order to obtain the morality and emotion reflected, it was employed a **keyword search** approach between each lexicon and the corresponding texts to reflect the views, and a **'prompting'** approach to indicate this new information to the model. This involves providing the model a prompt to guide its behaviour, allowing the model to focus on specific aspects of the text, such as moral values and emotions, while performing the classification.

To determine whether the prompt method was effective in improving text comprehension, four different templates were evaluated, varying in complexity and amount of subjective information added. This also allowed to analyse how these variables influence the perception of the template and its information processing.

Below, there are the different templates used, where P is Prompt, $\{\dots\}$ is the original input and $\{ m_n\}$ and $\{ e_n \}$ are the moral values and the emotions, respectively:

Using only **MoralStrength**

- **P1**:‘{ m_1} : {. . . } ’
- **P2**:‘The text {. . . } reflects the moral value { m_1}’
- **P3**:‘The moral value { m_1} is reflected in the text: {. . . }’
- **P4**:‘The text {. . . } reflects varying intensities of morality such as: { m_1, m_2. . . m_n }’

Using only **DepecheMood**

- **P1**:‘{ e_1} : {. . . }’
- **P2**:‘The text {. . . } reflects the emotion { e_1}’
- **P3**:‘The emotion { e_1} is reflected in the text: {. . . } ’
- **P4**:‘The text {. . . } reflects different emotions such as: { e_1, e_2, . . . m_n }’

Using **both**.

- **P1**:‘{ m_1} and { e_1} : {. . . }’
- **P2**:‘The text {. . . } reflects the moral value { m_1} and the emotion { e_1}’
- **P3**:‘The moral value { m_1} and the emotion { e_1} are reflected in the text: {. . . } ’
- **P4**:‘The text {. . . }reflects varying intensities of morality such as: { m_1, m_2, . . . m_n } and different emotions such as: { e_1, e_2, . . . e_n } ’

In Figure 4.2, there is an example from ‘ELECTION’ dataset along with his label and the different templates.

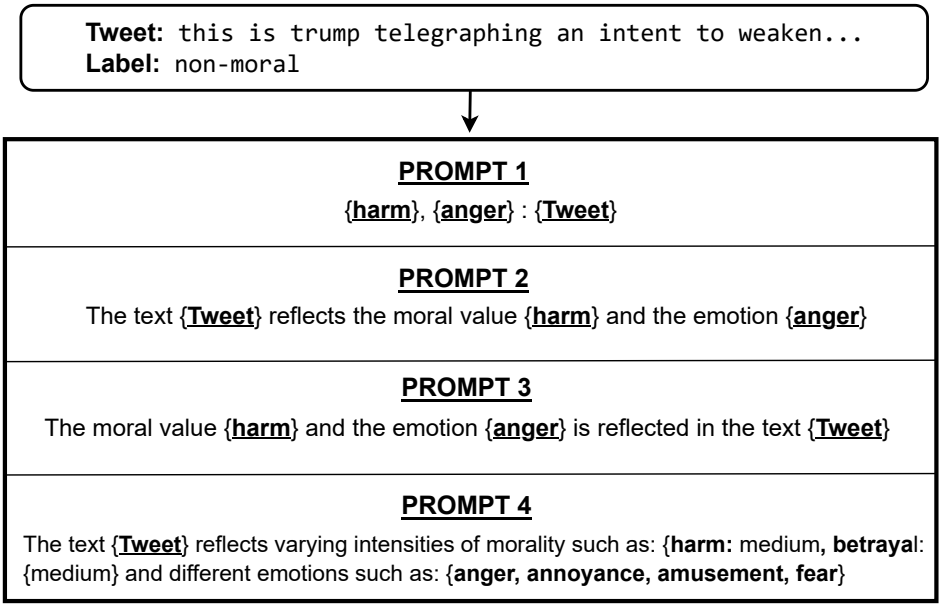

Figure 4.2: An example of the ELECTION dataset applied to the different templates and lexicons.

CHAPTER 5

Methodology

This section describes the resources used in this project, including datasets and lexicons, as well as the model and the metrics used to evaluate its performance. In addition, the experiments carried out are described in detail.

5.1 Model

The use of pre-trained language models has been explored, specifically **BERT base uncased** [25] and **RoBERTa base** [63], acquired from the Hugging Face framework [6].

Both models, belonging to the BERT family of models, were used for fine-tuning and inference use identical training parameters: **15 epochs**, a **learning rate** of **2e-5**, a **weight decay** of **0.01**, and a **batch size** of **32**.

Alongside the model, their correspondent tokenizer is used to transform raw data into a token sequence, that is the basic unit accepted for transformer models. Finally, fine-tuning process was performed, this consists of training the pre-trained base model with a specific data not seen for the model.

5.2 Datasets

The use of a variety of thematically diverse data has been shown to improve the model's performance in classification tasks and in learning to generalise, as seen in [50]. For this reason, in this work, use has been made of the **Moral Foundation Twitter Corpus (MFTC)** [38] and the **Moral Foundation Reddit Corpus (MFRC)** [89]. The MFTC is a dataset consisting of 35,108 tweets divided into seven thematically diverse subsets, ranging from social movements to climate events, which have been annotated by distinguishing between the ten categories defined in the Moral Foundations Theory (MFT), plus the label of non-morality.

The dataset is composed of the following subsets.

- **ALM**: tweets related with a social movement All Lives Matter in USA associated to criticizing the Back Lives Matter movement
- **BLM**: tweets related with a social movement dedicated to fighting racism and police brutality in USA Back Lives Matter
- **BALTIMORE**: referencing the Baltimore protests in the USA following the death of a young African American man at the hands of the police
- **DAVIDSON**: contains texts on hate speech and offensive language collected by Davidson et al.'s [23].
- **ELECTION**: tweets about the 2016 US presidential election.
- **SANDY**: contains messages about Hurricane Sandy that hit the US in 2012

This set of human-annotated English tweets has labels of moral foundations in 10 classes distinguishing between vice and virtue for each moral trait, including a 'non-moral' class. Tweets were tagged following the MFT and each domain was evaluated by at least three trained annotators as set out in the original labelling guide [39], which has been designed as a comprehensive manual that establishes common practices and clear guidelines for the identification of moral sentiments expressed in texts. Each tweet was therefore labelled with an indication of the presence or absence of each virtue and vice or using a 'non-moral' label.

The Reddit dataset consists of approximately 16,000 English-language comments drawn from the Reddit online platform, from 12 different subreddits. These subreddits are focused on specific topics and have been annotated according to the MFT, without distinguishing between the polarity of each of the moral dimensions. For this reason, the experiments

carried out with this dataset have focused on detecting the presence of a specific moral foundation and morality in general (MPres and MultiPres). Every post in the MFRC has been labelled by at least three trained annotators from a set of five for 8 categories of moral sentiment as outlined in the new version of the annotation manual, described in [89].

The label set for each tweet reflects the aggregated annotation of multiple annotators, using the majority vote as the true label. Table 5.1 shows the distribution of the different morals used, with a notable **imbalance** in which the non-moral class predominates. Differences of as much as 11,000 texts are found between the majority and minority class in datasets such as Reddit, and 2,666 texts in the Davidson dataset.

Dataset	C/H	F/C	L/B	A/S	P/D	NM
ALM	1,314	723	408	274	182	585
BLM	1,048	934	528	491	253	1,040
BAL	434	292	895	120	37	2,366
DAV	447	130	319	1,039	118	2,784
ELE	798	736	286	177	349	2,019
RED	2,240	1,784	610	1,204	435	11,435
SAN	708	708	1,010	519	560	291

Table 5.1: Class distributions of MFTC and MFRC. The moral foundations are described as C/H: care/harm, F/C: fairness/cheating, L/B: loyalty/betrayal, A/S: authority/subversion, P/D: purity/subversion and NM: non-moral.

5.3 Lexicon approach

The lexicons used **reflect** two types of **subjective perspectives**: **morality** and **emotions**. As shown in Figure 5.1, to assign one or more morals to each text, the **MoralStrength** framework was used, which provides functions to estimate the morals present in each text. These estimates reflect the polarity of each moral trait, with values ranging from 0, indicating virtue, to 9, indicating vice. In the case of prompts with only one moral, the difference between the extremes (0 or 9) and the neutral value (5) was calculated for each morality. The text was associated with the morality with the largest difference between these extremes. For prompts indicating several morals and their intensity, each difference was assigned a value (low, medium or high) according to the size of differences obtained. If the difference was greater than 2.5, it was classified as high; if it was between 1.5 and 2.5, it was considered moderate; and if it was less than 1.5, it was classified as low.

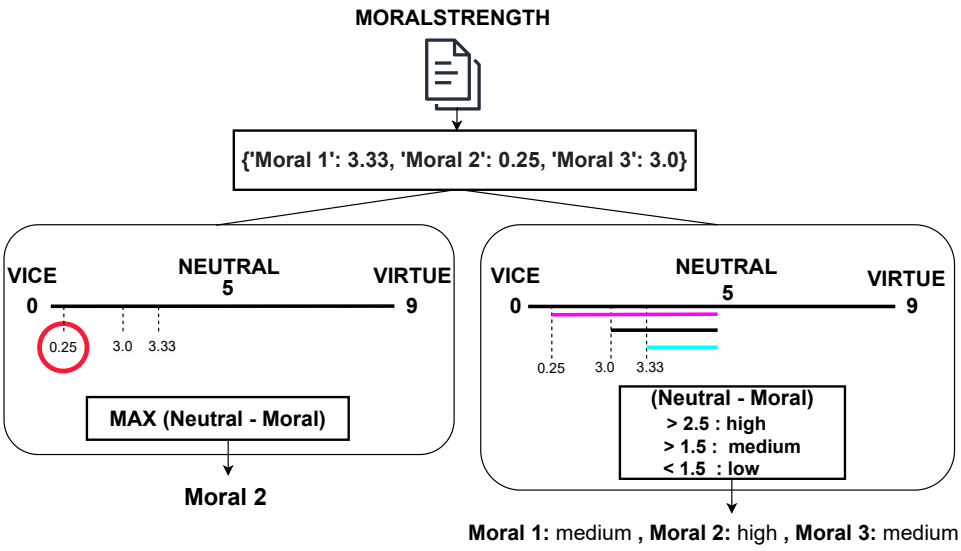

Figure 5.1: MoralStrength Lexicon approach. Lexicon used to select one (left) or multiple morals (right) based on the difference with the neutral value.

To associate each text with one or more emotions, a method of matching keywords present in the **DepecheMood**(++) lexicon was used. As shown in Figure 5.2, the lemmas in the lexicon were used, counting the presence of the lemmas in each text. The mean of the emotion scores of the matching lemmas in the text was then calculated. In case of a single emotion associated with the text, the emotion with the highest probability was selected. In situations where multiple emotions were associated with the text, the for with the highest

probability were identified as the most representatives.

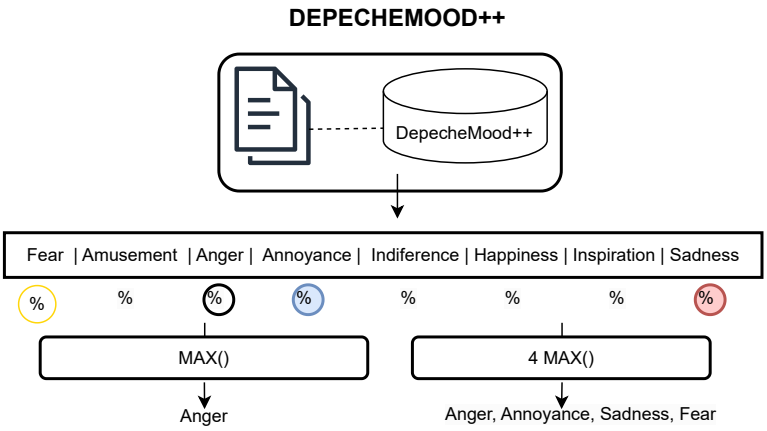

Figure 5.2: DepecheMood(++) Lexicon approach. Mean emotion probability is calculated, and the highest (black circle) is selected to represent one emotion, or the highest (black, blue, red circles) to represent several emotions.

5.4 Metrics

To assess the performance of the model, the **macro F1-Score** metric (5.3) has been chosen because of its wide adoption in similar work and its usefulness in situations of class imbalance, as it assigns a balanced weight to each class, regardless of its frequency.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.2}$$

$$F_1Score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5.3}$$

In order to evaluate and compare more precisely the results obtained for each model using different prompts and lexicons, and to determine whether there are significant differences in the improvement (allowing to compare them with previous work), the strategy followed in [12] has been replicated and the **Friedman test** has been applied. This test, is a non-parametric statistical tool used to determine the significance of differences between

multiple classification algorithms evaluated on multiple data sets. It provides a ranking of the classifiers in order of performance.

The formula for calculating the Friedman statistic (X_f^2), is defined as:

$$X_f^2 = \frac{12n}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{5.4}$$

Where n is the number of observations (sample size), k is the number of classifiers (or models) being compared R_j is the sum of the ranks assigned to classifier j, calculated from the rankings obtained on each data set. After calculating the Friedman statistic, it is compared to the chi-square distribution with k-1 degrees of freedom. If the calculated value of X_f^2 is greater than the critical value of the chi-square distribution for a defined level of significance ($\alpha = 0.05$), then it is considered that there are significant differences between the tested classifiers.

While the Friedman test allows for the comparison of multiple classifiers, an additional statistical test is required to compare pairs of classifiers. For this purpose, the **t-test** is employed and defined as:

$$t = \frac{\overline{d}}{s_d/\sqrt{N}} \tag{5.5}$$

Where t is the statistic, $\overline{d}$ is the mean of the differences between the performance of the classifiers, s_d is the standard deviation of the differences between the scores and N is the number of observations or pairs of classifications compared. This compares the performance means of two raters and determines whether there is a significant difference between them. To assess statistical significance, the p-value representing the probability of obtaining the observed results if the null hypothesis (H_0)(that there is no difference between the raters) is true is calculated. If the calculated p-value is less than the desired significance level ($\alpha = 0.05$), then the null hypothesis is rejected, it is concluded that there is a significant difference between the classifiers, and there is sufficient evidence to claim that the observed difference is not due to chance.

5.5 Experimentation

5.5.1 Data Pre-Processing

In the first phase of data pre-processing, the raw data extracted from both the MFTC and MFRC datasets were thoroughly cleaned to remove elements that could potentially interfere with the model's learning process. This involved a number of essential steps, including r**emoving emoticons, special characters and normalising text by converting it to lower case. URLs, usernames and hashtags were also removed** to ensure the integrity and coherence of the data. Despite the considerable length of the text within the datasets, it was decided not to remove stop words, as these could contain valuable information for the understanding of the model. Once the data was cleaned, each dataset was split into separate training and evaluation sets, allowing an independent assessment of the model's performance on different subsets of the data.

After the data pre-processing, the fine-tuning process was started by adding a fully connected layer on top of the pre-trained model encoder. This additional layer allowed the model to be adapted to the specific needs of our moral trait detection tasks. Setting the number of neurons in the fully connected layer to match the number of classes in each task ensured that the pre-trained model could adapt and understand the unique nuances of our datasets. The **data was then split into training and test sets using an 80-20% ratio**.

5.5.2 Fine Tuning

After splitting the data into each domain, **fine-tuning experiments were carried out independently for each dataset**. This approach allowed the model to be adapted to the specific characteristics of each domain, thus optimising its ability to detect moral traits in texts within that particular context.

In a first fine-tuning experiment, the pre-trained model was used together with its corresponding tokenizer 5.1. The aim was to obtain the baseline results of the tasks without the addition of lexicons, and to be able to compare the impact after their addition. Given the class imbalance seen in Table 5.1, which could affect the learning of the model, **different data balancing methods were applied**, such as oversampling to match the number of examples to the majority moral class, and undersampling to match the number of examples to the minority moral class of the task. After comparing the approaches, the most suitable approach was selected to continue training the model.

Once the models were trained, weights were loaded to make inferences. A 'Softmax' layer was used to convert the model outputs into probability distributions, allowing the model to detect moral foundations (MPres), determine the polarity of these foundations (MPol), assess one moral foundation at a time (MultiPres) or assess the overall moral polarity of the text (MultiPol), with the predicted class being the one with the highest probability.

For moral rationale detection **(MPres)**, a **binary classification task (moral or non-moral)** was performed. The polarity distinction of each moral foundation **(MPol)** was a **multi-class classification with 3 classes (non-moral, vice, virtue)**. The identification of the moral trait **(MultiPres)** was a **multi-class classification task with 5 classes (the 5 moral foundations and a non-moral class)**. Finally, the task of distinguishing polarity between moral traits **(MultiPol)** involved a **multi-class classification with 11 classes** (each moral trait with its corresponding **vice or virtue extreme and a non-moral class)**.

To incorporate the different lexicons using the **prompting approach**, a similar procedure was followed, **adapting the input data to the different templates** (4.2). We assessed how the inference results varied by looking at the macro F1 score, which allowed us to determine the impact of the lexicons on the classification tasks. Figure 5.3 shows an overview of different approaches to text classification.

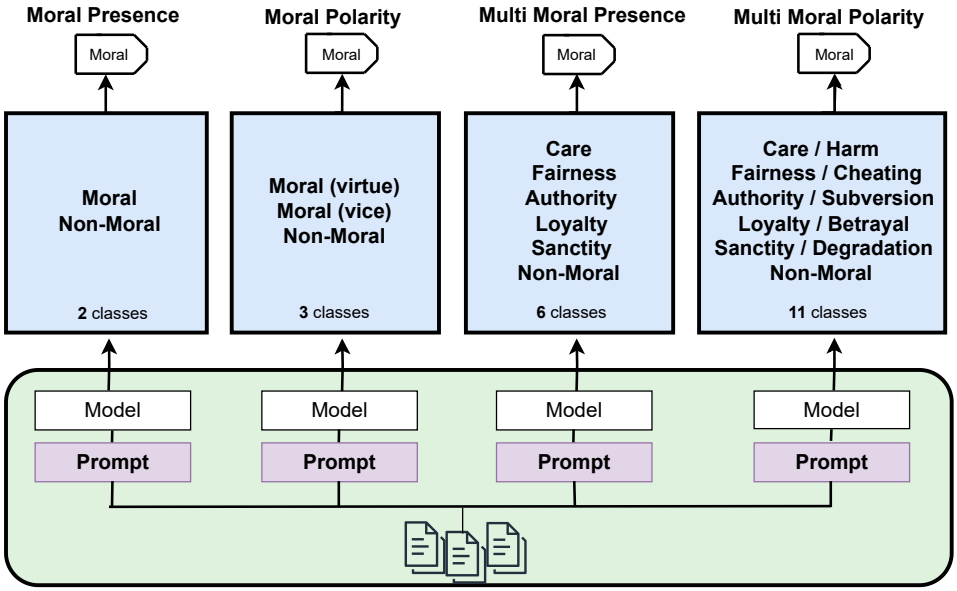

Figure 5.3: Fine Tune Model for tasks defined.

5.5.3 Generalization Across Domains

Given the importance of understanding language in different contexts, we explore how incorporating lexical information could improve the predictive ability of the model in different domains. After identifying the best indicator and lexical usage using the indicator method, an additional experiment was conducted to investigate its impact. By transferring the knowledge learned from the model during the fine-tuning process between datasets, the ability of the model to adapt to new domains was evaluated, trying to capture the nuances and complexities present in different textual contexts.

To **analyse model performance in varied domains**, a **cross-dataset validation process** of all datasets was carried out. **Models were trained using each dataset individually and evaluated through cross-testing between the remaining datasets**. Once the results were obtained, the corresponding metrics were evaluated with the baseline without lexicons.

Knowledge transfer experiments were conducted between datasets to assess the model's ability to infer in unseen domains. The fine-tuning technique was used on a specific domain and its performance was evaluated when making inferences on data from unseen domains, comparing the predictions with the original labels.

These experiments provided valuable information on the generalisation and adaptation capabilities of the model, which are fundamental for developing robust models capable of handling the diversity and complexity of language in different contexts.

5.5.4 Diverse Annotator Perspectives

This additional experiment focuses on incorporating different annotators' perspectives to improve the model's ability to estimate moral values in texts. This is done by **training language models with the annotations of different annotators and then combining these models in an ensemble approach**. In addition, the **evaluation of disagreement estimation in texts**, presented annotation challenges, will be conducted.

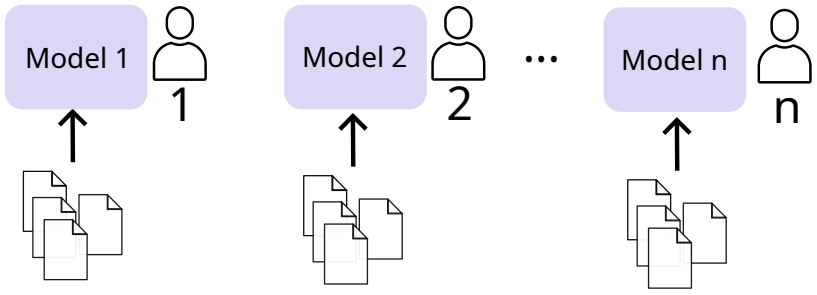

Figure 5.4: Fine-tuning procedure where models are trained with the specific annotations of n different annotators.

- **Annotation Diversity Exploitation:** In the process of annotator diversity, two main activities were conducted. First, language models were trained using the specific annotations provided by each annotator, as seen in Figure 5.4. This involved developing multiple models, each capturing the unique perspective of a particular annotator. Subsequently, a comprehensive evaluation of the classification performance of each of these models was carried out. This analysis allowed for an understanding of the consistency and variability in the annotations made by different annotators, providing valuable insights into how these discrepancies could influence the process of morality estimation in texts. In the ensemble approach, a method combining the predictions of individual models was employed. This strategy allowed for harnessing the diverse perspectives provided by different annotators, thus integrating a broader range of information into the process of morality estimation in texts. Additionally, a prompt-based approach was implemented to enrich the dataset with model predictions and conduct a second phase of training (Figure 5.5). This addition allowed the model to become even more familiar with the diverse perspectives and contexts present in the data, thereby enhancing its ability to understand and classify morality in different situations and topics.

- **Disagreement estimation** To evaluate the difficulty of annotating an instance, a

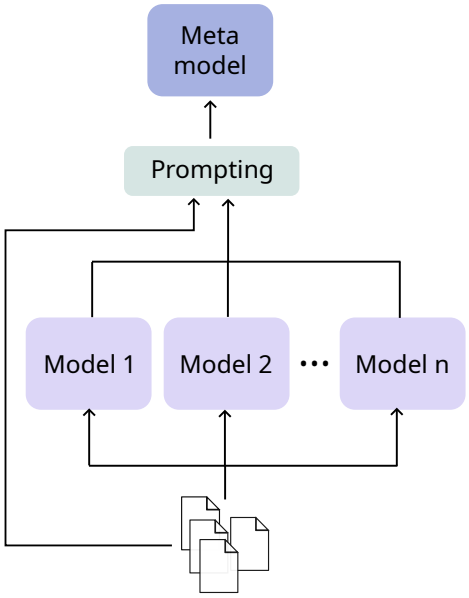

Figure 5.5: Proposed ensemble method that combined the predictions on the different perspectives of the base models with the textual input through a prompt approach.

threshold based on the divergence metric was defined. An instance was considered challenging to annotate if its divergence metric exceeded this threshold. Then, the SHapley Additive exPlanations (SHAP) method [66] was used to analyze how linguistic features affected the models' decision on disagreement in a document. Such a method assigns an importance score to each of the features considered for a specific prediction. These SHAP values allow to inspect the learning models trained, inspecting how the language affects the decision on the disagreement of a document.This evaluation provided insights into how language influenced the discrepancy between annotations and helped better understand the challenges in annotating moral texts. To perform this analysis, SHAP values were extract of all models trained, aggregating them to obtain a whole overview of the classification process. To do so, extracting the SHAP values for all words in all documents, aggregating them into a set of values for each word considered.

CHAPTER 6

Results

This chapter presents results of the experiments and statistical comparisons for evaluation.

6.1 Baselines

To obtain baseline results for comparison in each of the prediction tasks, fine-tuning was applied in three different scenarios: the original **imbalanced dataset (imb)**, the dataset balanced by **oversampling (over)**, and balanced by **undersampling (under)**. Table 6.1 shows the performance in terms of macro F1-Score obtained for each of the prediction tasks, reflecting each dataset in columns.

It can be seen that data imbalance significantly affects the model performance. Although RoBERTa shows a slight advantage over BERT in this case, the difference is not always significant and **results vary by domain**. For example, in the case of imbalanced datasets, BLM F1-Score values exceed 90.00%, while in DAV values are as low as 49.00%

Oversampling has a positive impact on performance, yielding almost optimal and similar results for both models, suggesting that both benefit equally from this technique. **When applying undersampling, similar or even better results compared to the imbal-**

anced case are obtained.

Although the results show that oversampling achieves very high F1-Score values, this strategy is not considered the best because it simply replicates data sets, which can lead to issues such as overfitting and loss of data diversity. Therefore, **the rest of the work has used data balanced with undersampling as the baseline results**, against which the prompting approach will be compared.

In the MPress task, although not remarkable, BERT surpasses RoBERTa in 5 of the 7 domains. In the MultiPres task, RoBERTa generally outperforms BERT, with more pronounced differences in domains like ALM (+4.51) and smaller differences in BLM (+0.25). However, there are exceptions, such as in the BAL dataset, where BERT outperforms RoBERTa with an F1-Score difference of 8.85%. Polarity tasks show similar performance in both models. **Overall, RoBERTa handles undersampling better compared to BERT**.

			ALM	**BLM**	**BAL**	**DAV**	**ELE**	**RED**	**SAN**
MPres	imb	BERT	79.61	91.81	**64.98**	**49.21**	75.45	66.24	75.72
		RoBERTa	**81.78**	**92.56**	62.22	49.14	**75.95**	**67.32**	**77.82**
	over	BERT	94.26	**98.06**	**96.42**	**95.72**	**97.36**	**97.55**	94.26
		RoBERTa	**94.47**	96.95	96.35	94.77	97.06	97.35	**94.42**
	under	BERT	78.79	88.56	**71.19**	**49.87**	79.17	73.10	78.12
		RoBERTa	**80.68**	**90.63**	63.95	48.34	**80.02**	**76.38**	**78.92**
MPol	imb	BERT	66.58	87.80	51.49	**32.97**	**62.57**	-	**60.48**
		RoBERTa	**70.26**	**88.67**	**51.83**	32.84	60.78	-	60.32
	over	BERT	**96.62**	**97.66**	**98.35**	**98.59**	**98.22**	-	**96.77**
		RoBERTa	96.07	97.32	97.57	98.37	98.14	-	96.10
	under	BERT	67.22	82.00	**56.26**	**31.79**	59.43	-	60.04
		RoBERTa	**73.56**	**86.03**	51.11	16.21	**69.46**	-	**61.06**
MultiPres	imb	BERT	61.20	85.13	37.23	15.79	54.27	46.35	55.57
		RoBERTa	**61.57**	**85.89**	**39.73**	**16.56**	**56.00**	**48.59**	**58.85**
	over	BERT	89.25	**87.67**	93.09	**92.23**	92.30	**96.28**	81.16
		RoBERTa	**89.42**	87.42	**93.57**	90.77	**92.58**	96.23	**82.07**
	under	BERT	60.66	81.21	**36.21**	15.72	62.98	49.20	**53.66**
		RoBERTa	**65.17**	**81.46**	27.36	**19.47**	**66.76**	**50.52**	53.06
MultiPol	imb	BERT	55.09	81.70	**29.72**	**8.65**	46.52	-	43.60
		RoBERTa	**56.21**	**83.01**	29.20	8.32	**47.24**	-	**48.90**
	over	BERT	**90.39**	**90.82**	94.04	91.89	**92.91**	-	83.61
		RoBERTa	90.38	89.90	**94.13**	**92.44**	**92.91**	-	**84.54**
	under	BERT	**64.80**	79.06	22.12	**14.25**	62.35	-	**55.09**
		RoBERTa	62.57	**80.70**	**27.55**	12.07	**62.65**	-	53.65

Table 6.1: Model comparison in terms of F1 scores in morality detection over different data balancing strategies.

There is a clear **decrease in F1-Scores as the tasks become more complex, and specific domains significantly influence the results**: domains like BLM consistently show high scores, while domains like DAV present considerable challenges for both models across the tasks.

For a comprehensive analysis of the models' performance, the Friedman ranking has been used. The average rankings of the classifiers reflect how each model performs relative to the other in terms of its ability on different tasks. A lower average rank indicates that a model has been consistently ranked higher, suggesting better overall performance; while a higher average rank indicates lower performance.

Looking at Table 6.2, the numbers in the Ranking' column represent the relative average position of each model on the different morality prediction tasks, with decimal values allowing raters to be compared in case of a tie. When **comparing the pre-trained models** against each other **in one task**, **there are generally no significant differences between them**. **Comparing between tasks, RoBERTa tends to outperform BERT in less complex cases**, such as MPres and MPol. As task complexity increases (as in MultiPres and MultiPol), the difference in performance between RoBERTa and BERT becomes less pronounced, even reaching a tie in the most complex task.

When analysing the performance of both models in relation to task complexity, a clear trend can be seen: as task complexity increases, general performance tends to decrease, suggesting that both **models** may **face additional challenges when considering more moral aspects**.

Friedman Test	**Ranking**
RoBERTa MPres	1.3
BERT MPres	1.7
RoBERTa MPol	3.9
BERT MPol	4.6
RoBERTa MultiPres	5.1
BERT MultiPres	5.7
BERT MultiPol	6.8
RoBERTa MultiPol	6.8
Statistics	
Chi2:	35.90
Friedman's F	16.45
F(7,42)—0.05	02.24

Table 6.2: Model Performance Ranking in different tasks.

With a significance level set at 0.05 ($\alpha = 0.05$), the calculated Friedman statistic was approximately 35.90, with a critical value of about 2.24 for 7 degrees of freedom. Because the value of the **Friedman statistic is much larger than the critical value**, the probability of obtaining a statistic value as large or larger under the null hypothesis H_0 (which states there are no significant differences between performance in tasks) is low. It is therefore rejected, and **it is concluded that there are more difficulties in models for detecting morality in more complex tasks.**

6.2 Prompting

The results presented in Tables 6.3 and 6.4 represent the average macro F1-Score of the various ranking tasks, with the purpose of comparing the performance of different prompting approaches. The aim is to investigate whether prompts containing information about morality with the **MoralStrength** lexicon **(MS)**, emotions with **DepecheMood (DM++)** or **both (MS/DM++)** affect the effectiveness of the models in classifying the tasks mentioned in 4.1. **Scores above the baselines are marked with an asterisk (*)** to analyse whether the use of a lexicon improves performance, focusing on the effectiveness of prompts rather than the lexicons themselves. Additionally, **the highest score for each dataset is highlighted to show the influence across domains**. The tables presented are averages across the classification tasks, designed to assess how different prompt approaches perform with various lexicons. This includes whether information about morality, emotions, or both is used in the prompts. Table 6.3 shows the BERT model results.

		ALM	BLM	BAL	DAV	ELE	RED	SAN	Fm
	Baseline	67.87	82.71	46.47	27.91	65.98	61.15	61.73	7.6
P1	MS	67.79	83.77*	43.90	26.07	68.76*	60.23	63.18*	9.3
	DM++	67.96*	82.20	44.93	25.27	67.57*	60.13	61.71	10.7
	MS/DM++	68.46*	83.45*	**49.21***	28.03*	67.79*	61.61*	62.30*	6.3
P2	MS	68.63*	83.92*	46.64*	24.86	69.33*	61.86*	61.22	7.0
	DM++	67.98*	84.25*	44.27	27.49	66.71*	62.20*	62.95*	6.3
	MS/DM++	68.72*	84.06*	47.46*	**29.14***	69.49*	61.15	64.25*	3.6
P3	MS	68.69*	**84.59***	44.53	26.55	69.11*	62.04*	63.42*	4.9
	DM++	67.83	83.49*	45.68	26.13	65.66	62.15*	**65.32***	7.6
	MS/DM++	67.89*	84.08*	45.24	27.35	**69.61***	61.17*	62.52*	6.4
P4	MS	**69.03***	84.23*	40.28	28.62*	67.72*	61.04	62.98*	6.6
	DM++	67.95*	83.77*	49.15*	25.88	68.13*	61.37*	63.52*	6.9
	MS/DM++	67.79*	84.22*	41.36	26.87	68.60*	**63.13***	63.68*	6.4

Table 6.3: Prompting results BERT. Pi represents Prompti where i=1,2,3,4.

Looking at the **BERT** F1-Score values, **simpler prompts perform better** (P1 and P2). The simplest prompt, P1, combined with MS/DM(++) stands out as performing best on most metrics, especially BAL and DAV with an improvement of 2.74% and 1.81% respectively. Similar results are observed in P2 with MS/DM(++), suggesting **that the combination of moral and emotional information is highly effective, even with less information provided**.

Regarding P3 and P4, although no specific lexicon stands out, some improvement are shown, such as using again both lexicons in the ELE dataset with (+3.63), a particularly high increase. In both cases, indicating the information before the text (P3) or providing more information in the input (P4) does not necessarily translate into a substantial improvement.

In general, all forms of adding additional information improve the results for BERT. Looking at Friedman's ranking (column Fm). 0n the one hand, it is noticeable that prompt 2 consistently outperforms other prompts in this analysis, obtaining the highest position (3.6). The other uses of the lexicons also achieve outstanding positions in the ranking. Moreover, it is the only case where the baseline is ranked lower in all lexicon use, with a ranking of 7.6 compared to 7.0 and 6.3. On the other hand, prompt 4 also outperforms the

baseline, although the difference between the different lexicons used is not as marked.

In this case, prompt 2 stands out as an effective option for improving the model's performance in moral classification tasks, according to Friedman's statistical analysis.

		ALM	**BLM**	**BAL**	**DAV**	**ELE**	**RED**	**SAN**	**Fm**
	Baseline	70.49	84.71	42.49	24.03	69.72	63.45	61.67	11.6
P1	MS	71.41*	**85.95***	46.32*	24.46*	71.20*	64.32*	**63.48***	4.6
	DM++	71.45*	84.88*	44.74*	24.38*	68.92	64.36*	63.20*	8.1
	MS/DM++	71.11*	85.37*	45.62*	24.06*	70.79*	64.26*	63.21*	6.9
P2	MS	71.93*	85.90*	44.93*	22.96	71.19*	63.87*	61.92*	7.9
	DM++	70.51*	84.95*	49.07*	26.04*	69.40	**65.36***	63.27*	5.7
	MS/DM++	**73.57***	85.60*	45.55*	24.09*	70.69*	64.30*	63.01*	5.8
P3	MS	71.55*	85.23*	45.85*	**26.73***	**72.28***	63.25	63.10*	6.0
	DM++	71.04*	84.94*	46.47*	23.58	70.77*	64.33*	63.00*	7.9
	MS/DM++	71.76*	84.94*	46.90*	21.12	71.84*	64.13*	61.83*	7.9
P4	MS	71.77*	85.19*	46.47*	23.14	71.30*	65.10*	62.98*	6.0
	DM++	72.42*	85.23*	**50.02***	24.69*	69.72	64.13*	62.74 *	5.9
	MS/DM++	71.70*	85.55*	48.72*	23.02	71.66*	64.15*	61.63	6.9

Table 6.4: Prompt results RoBERTa. P represents Prompti where i=1,2,3,4.

Table 6.4 shows the RoBERTa model results. In general, **it is observed an improvement in the performance of the RoBERTa model compared to the baselines in all domains, with even higher values in terms of F1 score compared to BERT.** Unlike BERT, there is no concrete pattern, but it is noteworthy that the use of a single morality lexicon gives better results compared to the baseline, with high improvements observed, as in the case of the Baltimore dataset with an improvement of almost 4.00%. **It should be noted that high results are obtained when the emotion lexicon is used instead of both, which is different from BERT who obtained them with both.**

With average rankings calculated, a significant improvement in the performance of the different prompts can be seen compared to the baseline (11.6). In particular, the case of **prompt 1 combined with MoralStrength stands out** with a mean ranking of 4.6, which reflects that it is in first place, indicating a better performance in general.

Comparing the two models, RoBERTa's metrics are equal to or better than BERT's in all prompt proposals (P1 to P4). It is interesting to note that in prompt 2, while for

RoBERTa the MS/DM(++) metrics are similar or slightly better than MS or DM(++) alone, for BERT the MS/DM(++) metrics are significantly better, showing that models are affected differently.

Different lexicons have a varying effect on model performance depending on the complexity of prompt used; in P1 and P2 performance is better with both (MS/DM(++)). However, in other prompts such as P3 and P4, there is no noticeable difference in performance between the different lexicons.

As results of these results, it is concluded that **simpler instructions perform better in the model performance**. Tables 6.7 and 6.8 show the F1-Score improvement in each domain (columns), together with the tasks differentiated from each other and ordered from least to most complex.

To conduct a more detailed analysis of the improvement using additional information, a **Friedman test was performed using the top three model in Friedman rankings and the baselines** to rank the classifiers based on their performance. A **t-test was then used to compare the top classifiers against the baseline**. This was done because the Friedman test requires at least three classifiers for proper application, and it was necessary to specifically compare the top performing classifiers to the baseline.

Friedman Test	**Ranking**
Prompt2 + MS/DM++	1.6
Prompt3 + MS	2.3
Prompt1 + MS/DM++	2.4
BERT Baseline	3.7
Statistics	
Chi2:	10.03
Friedman's F	5.48
F(7,42)—0.05	3.16
t-test (p value)	0.014

Table 6.5: BERT prompts performance ranking.

Based on the results obtained in the Friedman test of BERT model (Table 6.5), it is clear that certain instruction configurations (lexicons used) significantly outperform the BERT baseline. The prompt 2 MS/DM(++) combination achieved the best ranking position with a difference of 2.1 with respect to the baseline.

Statistical analysis supports these results, indicating significant differences in performance between the instruction configurations and the baseline. The chi-square value of

10.03 and Friedman's F-value of 5.48 indicate a statistically significant difference in rankings between groups ($p < 0.05$). About t-test, with a p-value of 0.012 obtained from the t-test, conducted with an alpha level of 0.05, there is strong evidence to reject the null hypothesis.

These results strongly suggest that **incorporating prompts, particularly simple ones like Prompt2, along with both types of information (MS/DM(++)), enhances BERT's classification capabilities.**

Friedman Test	**Ranking**
Prompt1 + MS	1.6
Prompt2 + DM++	2.1
Prompt2 + MS/DM++	2.4
RoBERTa Baseline	3.9
Statistics	
Chi2:	11.91
Friedman's F	7.9
F(7,42)—0.05	3.9
t-test (p value)	0.012

Table 6.6: RoBERTa prompts performance ranking.

In Table 6.6, RoBERTa model shows that the combination of prompt 1 and MS achieved the best ranking position, supported by statistical evidence. The chi-square value of 11.91 and Friedman's F-value of 7.9 indicate significant differences in performance among the prompt configurations and the baseline ($p < 0.05$). Furthermore, the t-test provides additional support, with a lower p-value indicating that **prompt 1 with moral information outperforms the RoBERTa's baseline.**

			ALM	BLM	BAL	DAV	ELE	RED	SAN
MPres		Baseline	78.79	88.56	**71.19**	49.87	79.17	73.10	78.12
	P1	MS	**81.35**	88.40	64.29	47.94	79.99	73.48	79.60
		DM$_{++}$	78.43	**88.84**	67.04	49.80	**80.29**	73.11	77.42
		MS/DM$_{++}$	79.89	88.46	67.21	49.71	78.27	**74.26**	78.85
	P2	MS	**82.08**	**90.10**	69.70	49.08	77.88	73.70	**80.07**
		DM$_{++}$	81.16	89.32	66.99	50.78	**80.46**	**73.93**	78.98
		MS/DM$_{++}$	81.21	90.08	67.25	**50.84**	79.09	73.75	79.99
MPol		Baseline	67.22	82.00	56.26	31.79	59.43	-	60.04
	P1	MS	67.67	85.38	48.79	28.61	64.98	-	60.00
		DM$_{++}$	69.50	81.41	**58.81**	25.26	62.05	-	57.59
		MS/DM$_{++}$	**69.94**	**86.31**	56.45	26.26	**65.89**	-	**60.11**
	P2	MS	69.34	**86.92**	53.78	22.44	71.69	-	54.82
		DM$_{++}$	71.30	84.93	56.82	26.49	66.41	-	59.02
		MS/DM$_{++}$	**71.43**	86.53	**58.69**	**31.98**	**73.92**	-	**61.32**
MultiPres		Baseline	60.66	81.21	36.31	15.72	62.98	49.20	53.66
	P1	MS	61.40	**81.70**	29.98	16.00	**67.02**	47.00	57.87
		DM$_{++}$	62.24	79.58	28.05	13.20	66.06	47.16	**58.84**
		MS/DM$_{++}$	**63.25**	79.38	**38.21**	**21.55**	66.84	48.97	55.49
	P2	MS	63.43	79.06	35.26	17.01	**67.47**	50.02	55.45
		DM$_{++}$	60.78	81.06	28.68	**19.92**	60.68	**50.47**	57.86
		MS/DM$_{++}$	**63.94**	77.85	**43.62**	16.72	64.92	48.55	**58.57**
MultiPol		Baseline	**64.80**	79.06	22.12	14.25	62.35	-	55.09
	P1	MS	60.73	79.60	32.53	11.72	**63.03**	-	**55.24**
		DM$_{++}$	61.67	78.98	25.81	12.82	61.87	-	53.00
		MS/DM$_{++}$	60.75	79.67	**34.98**	**14.63**	60.16	-	54.76
	P2	MS	59.66	79.58	**27.82**	10.93	60.28	-	54.55
		DM$_{++}$	58.67	81.68	24.59	12.77	59.29	-	55.93
		MS/DM$_{++}$	58.29	**81.80**	20.27	**17.03**	60.03	-	**57.10**

Table 6.7: BERT performance on moral classification tasks with added knowledge with lexicons and prompts.

			ALM	BLM	BAL	DAV	ELE	RED	SAN
MPres		Baseline	80.68	90.62	63.95	48.34	80.02	76.39	78.92
	P1	MS	80.51	**91.57**	**69.75**	**49.50**	80.77	75.97	79.42
		DM++	80.52	91.07	65.21	48.24	**81.05**	76.03	**80.46**
		MS/DM++	**81.93**	90.69	**64.48**	48.47	80.15	75.58	79.58
	P2	MS	81.53	90.75	64.80	44.56	82.26	75.85	79.14
		DM++	80.86	91.29	64.50	49.35	81.38	**77.06**	79.24
		MS/DM++	**83.11**	**91.32**	**70.91**	**52.37**	**82.16**	76.75	**79.75**
MPol		Baseline	73.56	86.03	51.11	16.21	69.46	-	61.06
	P1	MS	71.46	**88.50**	53.85	18.45	**74.03**	-	**62.24**
		DM++	73.09	86.30	**56.56**	17.37	70.23	-	59.79
		MS/DM++	71.78	88.19	50.76	**20.99**	73.35	-	61.27
	P2	MS	73.69	**88.66**	51.76	**20.14**	**74.51**	-	57.93
		DM++	**74.33**	86.73	**52.76**	20.09	71.69	-	**62.76**
		MS/DM++	72.82	88.42	49.31	15.42	72.93	-	61.61
MultiPres		Baseline	65.17	81.46	27.36	19.47	**66.76**	50.52	53.06
	P1	MS	**68.33**	**82.86**	30.81	19.97	66.53	52.67	**56.40**
		DM++	66.34	80.90	24.78	**20.25**	62.63	52.68	55.99
		MS/DM++	66.99	80.77	**33.41**	14.78	66.08	**52.94**	55.31
	P2	MS	68.15	**82.69**	30.85	13.07	65.48	51.89	56.86
		DM++	64.07	81.78	**41.45**	16.15	63.08	**53.66**	**57.79**
		MS/DM++	**70.98**	81.92	29.87	13.61	63.53	51.84	56.06
MultiPol		Baseline	62.57	80.70	27.55	12.07	62.65	-	53.65
	P1	MS	65.32	80.89	30.85	9.91	63.46	-	55.87
		DM++	**65.86**	81.25	32.42	11.67	61.76	-	56.55
		MS/DM++	63.75	**81.82**	**33.81**	12.02	**63.58**	-	**56.68**
	P2	MS	64.34	**81.51**	32.32	14.07	62.51	-	53.76
		DM++	62.80	79.98	**37.57**	**18.57**	61.43	-	53.31
		MS/DM++	**67.38**	80.73	32.13	14.98	**64.14**	-	**54.61**

Table 6.8: RoBERTa performance on moral classification tasks with added knowledge with lexicons and prompts.

6.3 Cross Dataset

Figure 6.1 shows the distribution of the difference between the base-case performance and the use of different prompts (1,2 and 4) and lexicons (MS and MS/DM++)for the Morality pressence task. With the use of the MS lexicon, greater improvements are generally observed as the complexity of the prompts increases in a cross-domain environment. The box plots represent the interquartile range (the central 50% of the data), with the central line indicating the median improvement. For the most complex prompt, P4, a marked improvement is observed, with the median indicating an approximate 1.5% increase in moral presence detection, and an interquartile range of up to 2%, compared to simpler prompts where a greater presence of negative results ('deterioration') is observed.

Using both lexicons does not provide significant benefits and may even negatively impact performance, especially when the complexity of the prompt increases (MS/DM++). This may indicate that subjective information saturation does not effectively contribute to the improvement of moral presence detection. It highlights that the effectiveness of using different prompts in combination with a specific lexicon (such as MS) may vary depending on the application context.

Furthermore, it is noteworthy that prompt 4, which includes multiple moral values, shows superior effectiveness in cross-domain settings, while in non-cross-domain settings, prompt 1, which addresses only one moral value, performed better. This suggests that incorporating more than one moral value (as in prompt 4) can enhance performance in diverse scenarios, whereas focusing on a single moral value (as in prompt 1) might be more effective in a more consistent or homogeneous dataset.

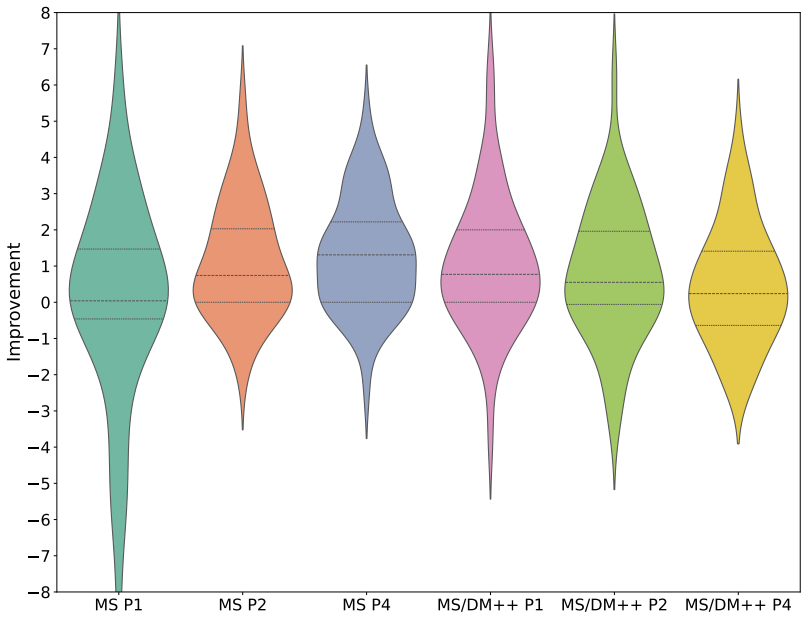

Figure 6.1: Comparison of moral presence detection using different prompts and lexicons, distribution of difference between baseline and lexicon approach

For more details on the most significant improvements, table 6.9 shows the percentage increases in moral presence detection using the MS lexicon with the most complex prompt (P4). The rows represent the domains used for training, while the columns represent the inference domains. The values show the percentage increase over the baseline.

The ELE domain shows significant improvements in all inference domains, especially when trained on the BLM (+5.14%) and DAV (+2.46%) datasets. Notably, training on the ALM domain produces the largest improvement in the SAN domain (+3.89%), highlighting its robust applicability across different contexts. The BAL domain shows consistent improvements in almost all inference domains except DAV. However, training in the RED domain sometimes has negative effects, especially in the ALM domain (-2.35%).

Using the MoralStrength lexicon with a complex stimulus generally improves performance in several domains, although some benefit more than others. It is important to carefully select combinations of training and inference domains to maximise recognition accuracy.

Figure 6.2 illustrates the difference in performance between the base case and various prompts and lexicons used for the Morality assessment task. Similar to the Moral Presence task, prompts 1, 2, and 4 were selected with the lexicons MS and MS/DM++. In this case, there is no significant difference between the use of one or both lexicons. However, in

	ALM	BLM	BAL	DAV	ELE	RED	SAN
ALM	-	+0.27	-0.76	-0.07	+1.31	+1.73	+3.89
BLM	+3.91	-	+1.64	-0.21	+3.24	+0.41	+1.27
BAL	+3.36	+3.28	-	-1.05	+3.90	+3.04	+3.04
DAV	+1.48	+1.71	+1.35	-	+1.73	-0.61	+1.06
ELE	+1.59	+5.14	+1.75	+2.46	-	+3.51	+2.22
RED	-2.35	+1.43	+2.38	-0.26	+1.79	-	+1.32
SAN	+0.72	+0.08	+0.62	+1.27	+1.05	0.00	-

Table 6.9: Impact of MS P4 on Baseline.

contrast to the detection of the presence of morality, in this case the use of both lexicons with the more complex task (P4) stands out, as it produces a considerable improvement. It is observed that both the maximum value of the interquartile range (boxplot box) and the median exceed all cases, reaching approximately 4% and 2%, respectively.

In this context the use of emotional values reflected allows the model to more fully capture the underlying morality. By considering a wide range of moral values, the model can more accurately assess the diversity and complexity of the text.

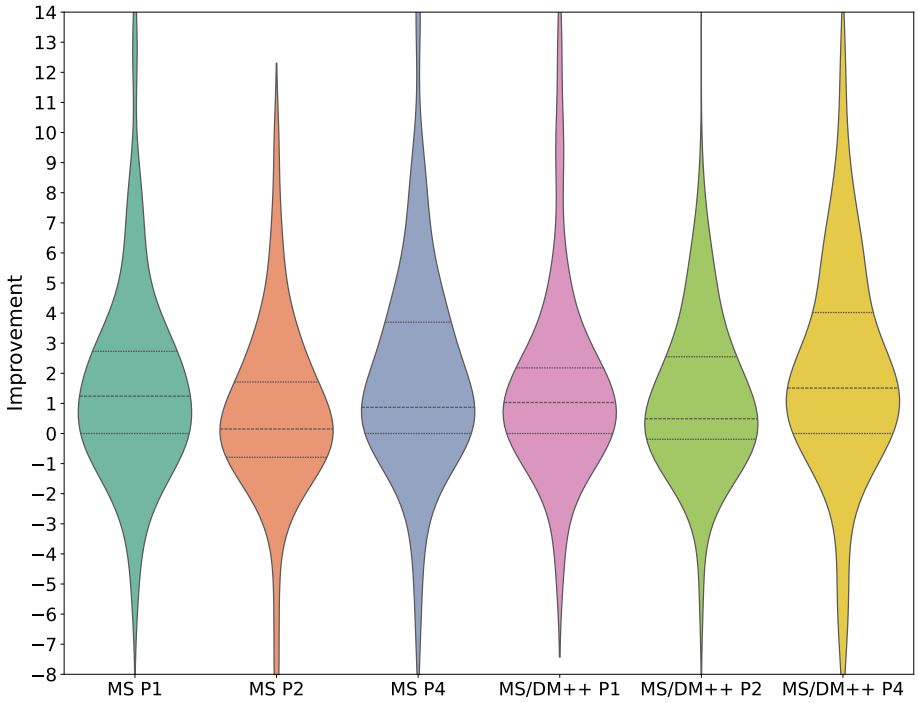

Figure 6.2: Comparison of moral presence detection using different prompts and lexicons, distribution of difference between baseline and lexicon approach.

As seen in Table 6.10, in general, the BLM and RED domains benefit the most from the combination of the MoralStrength and DepecheMood(++) lexicons with the more complex prompt, achieving the greatest improvements, reaching up to 12%. Other domains, such as ELE and SAN, also show improvements, but with greater variability and sensitivity to the training data. Notably, DAV does not produce the best inference results. This could result from several factors, such as the complexity of moral discourse in tweets or mismatches between lexicons and domain language. seen in Table 6.10, in general, the BLM and RED domains benefit the most from the combination of the MoralStrength and DepecheMood++ lexicons with the more complex prompt, achieving the greatest improvements, reaching up to 12%. Other domains, such as ELE and SAN, also show improvements, but with greater variability and sensitivity to the training data. Notably, DAV does not produce the best inference results. This could result from several factors, such as the complexity of moral discourse in tweets or mismatches between lexicons and domain language. **In general, the combination of MoralStrength and DepecheMood++ with the more complex prompt is a promising strategy for improving the detection of moral presence across a variety of domains.**

	ALM	**BLM**	**BAL**	**DAV**	**ELE**	**RED**	**SAN**
ALM	-	+9.44	+1.85	+1.56	+6.37	+5.38	+4.60
BLM	-4.11	-	-5.44	+2.47	+5.70	+6.45	+1.15
BAL	+1.29	+12.16	-	-0.34	+4.54	+1.37	+2.57
DAV	+5.38	+8.17	+0.05	-	+1.84	+1.92	+4.02
ELE	+1.51	+5.82	-1.07	-1.71	-	+1.08	-1.12
RED	+2.76	+7.40	+3.56	+1.15	+1.95	-	+1.26
SAN	-6.53	+0.47	+1.24	+0.32	+1.53	+2.28	-

Table 6.10: Impact of MS + DM P4 on Baseline.

6.4 Diverse Annotator Perspectives

In relation to the research questions 3 and 4; Table 6.11, shows in terms of the model's performance that when **using several perspectives in prompts, a significant improvement in the classification performance was obtained** in all domains. This suggests that the choice of prompt and additional information on different perspectives can influence and improve the results. The incorporation of this additional information has effectively provided more contextual cues, allowing the model to better understand and classify morality in different texts across various domains. Moreover, the observed improvements

in F1 scores highlight the effectiveness of leveraging diverse perspectives from annotators. By adding these into the training process, the model becomes more efficient at recognizing moral nuances present in texts. However, it's remarkable that while the prompt-based approach has led to considerable enhancements, certain domains, like Davidson, still present challenges for accurate classification.

	F1-score	
	Baseline	**Prompting**
ALM	64.71	88.74
BLM	85.46	95.82
Baltimore	42.58	76.32
Davidson	15.84	66.03
Election	61.11	88.22
Sandy	55.73	86.44

Table 6.11: Evaluation of the addition of different perspectives in training. The F1-Score results are compared with baseline results in all domains.

Analysing the nature of annotator disagreement by training a learning model to predict whether a given text presents annotation challenges, Table 6.12 reveals that, in some cases, the classifier effectively distinguishes divergent instances, particularly evident in the BLM, Baltimore, and Election domains, which exhibit the highest performance metrics. However, in domains like ALM, Davidson, and Sandy, classifiers struggle to discern divergence, albeit achieving an F-score of 58% for ALM and Sandy. This variance among domains is consistent with previous studies, indicating that differences in MFTC domains influence prediction task quality.

	Acc.	**F1-score**	**Neg. inst.**	**Pos. inst.**
ALM	58.55	58.36	94	99
BLM	68.94	68.49	120	115
BAL	79.26	79.25	402	355
DAV	48.17	47.75	442	403
ELE	71.61	71.39	272	288
SAN	58.82	58.81	180	177

Table 6.12: Evaluation in the task of predicting whether a text is challenging to annotate with morality. Accuracy, macro averaged F-score, and the number of negative and positive instances are reported.

Overall, these findings highlight the existence of language cues that indicate to learners whether a text is challenging to annotate. Given that these language cues may vary across

different annotation domains (and in addition with difference between domains adding lexicon information), a detailed analysis was done. SHAP method was used to analyse how learners interpret texts in terms of divergent annotations. Figure 6.3 shows the results obtained from a selection of the tokens that have the highest relevance for either the negative or positive classes. Tokens with negative SHAP values are relevant for detecting the negative class (low disagreement), while tokens with positive SHAP values are related to detecting the positive class, where the disagreement is higher. It is observed that tokens with negative SHAP values are generally words with semantics not pertinent to morality and innocuous in terms of societal or cultural issues; Interesting examples are photo, wonderful, green, internet or babies. This is an intuitive result since annotators will generally agree within texts that do not convey a strong moral or cultural position. In contrast, tokens with positive SHAP values tend to express strong moral significance. Some examples of these words are democrats, evil, god, racism, homo (from homosexuality), and respect. Again, this can be explained if we consider that annotators will disagree more frequently when assessing documents that include morally and culturally stronger positions. Interestingly, some tokens with higher positive SHAP values revolve around polemic or even harmful matters such as religion, sexual practices, and racism.

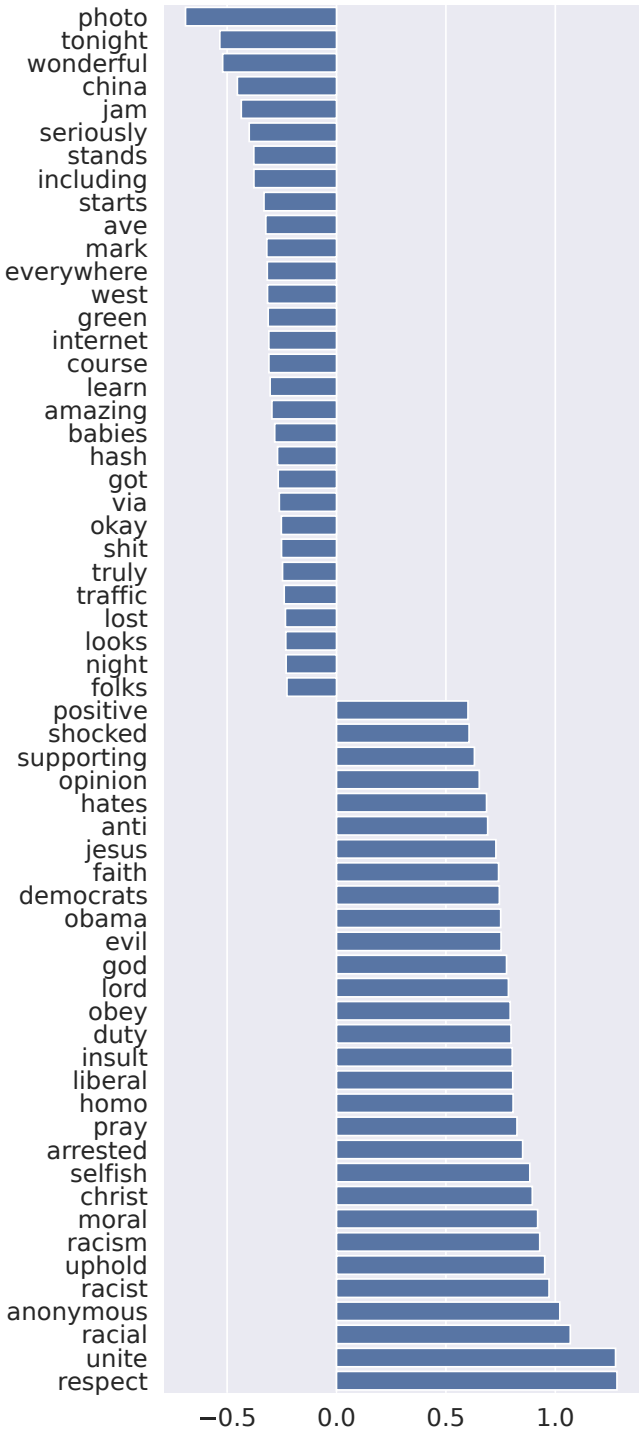

Figure 6.3: SHAP values of interesting tokens. Positive values indicate relevance towards the positive class, while negative values indicate otherwise.

CHAPTER 7

Conclusions and future work

In this chapter it will be describe the conclusions drawn from this project, and and the possible avenues for future work.

7.1 Conclusions

As the complexity of the task increases, there is a noticeable decline in the F1 scores of both models, where the results vary significantly across different domains and thematic areas. The task of detecting the presence of moral traits (MPres) is the least complex and thus has the highest scores. In contrast, tasks involving the evaluation of general morality (MultiPres) and the inclusion of his polarity in moral evaluation (MultiPol) are the most complex, showing the greatest declines in F1 scores.

When enriching the input with subjective information such as emotions and moral knowledge through prompting, differences emerge between the pre-trained models. Althought RoBERTa appeared to perform better across all; as task complexity increased (classes to classification), the performance gap between the models narrowed.

In analysing which type of information influenced the results, it was observed that BERT

had a greater impact when using the morality lexicon or both lexicons simultaneously. In contrast, RoBERTa showed a significantly greater impact when using the emotion lexicon. Although, in general, the addition of any knowledge to the input did not degrade performance, it is true that emotions had a more substantial influence on RoBERTa than on BERT. On the other hand, the way information was presented to the model also influenced performance. For example, BERT tended to benefit from more complete cues that included structured context along with the original text, whereas RoBERTa showed improvements even with simpler cues involving direct concatenation of lexical information.

When evaluating RoBERTa's generalization across different contexts, although expectedly the results were poorer than when fine-tuning on each specific domain, they were not drastically low. However, the effectiveness still depended on the domain or dataset type, indicating that certain messages or texts reflecting specific themes were more challenging for the model to evaluate. The method of adding knowledge through prompting also improved results across domains, aiding in transfer learning and enhancing the model's generalization capabilities.

7.2 Research Outcomes

In this section, the conclusions and answers to the proposed research questions are presented. The outcomes achieved are explained, and the improvements obtained are discussed

Firstly, **RQ1** inspects how varying levels of task complexity impact the accuracy and performance differences between two models used for moral prediction in text. This question has been addressed by evaluating the F1 scores of BERT and RoBERTa across tasks of differing complexity, showing a noticeable decline in F1 scores for both models as task complexity increased.For instance, the task of detecting the presence of moral traits (MPres) is the least complex and yielded the highest scores. In contrast, tasks involving the evaluation of general morality (MultiPres) and the inclusion of polarity in moral evaluation (MultiPol) were more complex and resulted in the greatest declines in F1 scores. This demonstrates that as task complexity increases, the efficiency of training and the effectiveness of predictions are negatively impacted, validating the hypothesis that task complexity significantly affects model performance. Secondly, **RQ2** explores the effect of integrating subjective information, such as emotions and moral knowledge embedded within text, on the generalization capacity of learning models used for automated moral assessment. This has been examined by enriching the input with subjective information through prompting and observing the performance differences between BERT and RoBERTa. The findings

indicate that RoBERTa performs better across all tasks when subjective information is included. Notably, as task complexity increased, the performance gap between the models narrowed. This suggests that incorporating subjective information generally enhances the models' ability to generalize moral evaluation, making them more robust and adaptable across different contexts.

In **RQ3** it was inspected the effect of exploiting the diversity of annotators' perspectives for automated moral estimation. It is shown that the different annotators do highly impact the quality of the predictions if taken in isolation. Attending to this, it is clear that the diversity of annotators and domains are variables to take into account when generating new data repositories. In contrast, the experiments show notable and consistent improvements in the classification performance when adding the predictions of models trained to estimate individual annotators' perspectives into an ensemble model. Such a positive result motivates future research on harnessing diverse perspectives into learning systems. Finally, **RQ4** proposes the task of estimating whether a data instance is challenging to annotate. That is, if an instance generates disagreement among annotators. Through this task, we intend to analyse the linguistic cues that indicate disagreement factors. The experiments show that the ability to estimate disagreement can achieve high performance scores but varies across domains, indicating considerable variance. By doing a subsequent analysis using SHAP values, we have discovered that the disagreement instances tend to contain strong moral, political, or cultural meanings. On the contrary, instances where annotators typically agree normally contain more neutral language.

7.3 Future work

Future research should address several limitations noted in this study, such as the data volume. Despite having a substantial dataset (over 35,000 texts), these models are typically trained on millions of data texts from diverse sources. Additionally, the inherently subjective nature of the task, as mentioned in Section **??**, implies that both the data annotations and the added subjective information can introduce biases, potentially hindering accurate classification. Future work will explore the use of multi-label prediction to account for the presence of multiple moral traits simultaneously, involving consideration of all possible moralities for the detection of multiple moral traits at once.

Furthermore, other methods of introducing subjective information will be explored, as well as other models like LLMs. This exploration may include the possibility of embedding subjective information directly or incorporating different types of data.

CHAPTER 8

Impact of this project

The project focuses on the development of a system for estimating moral foundations based on natural language processing techniques and transformational models. As mentioned in the introduction, in a world where digital platforms have radically altered our interactions and the way we consume information, there is a pressing need to understand how these channels influence our online perceptions and behaviours [65]. Moreover, the increasing emergence of polarised debates and divisive discourses in the digital sphere poses an urgent challenge to be solved by detecting [74, 94]. In this context, the project proposes to address these challenges by developing a system capable of estimating the moral foundations present in online discourse, in order to promote more respectful and understanding communication in social networks and other digital environments. This project not only addresses the challenges of online communication, but also has profound implications for various aspects of society and technology. These include improving the quality of public debate, promoting ethics in technology development and promoting a more inclusive and respectful digital environment.

8.1 Social, ethical, economic and environmental impacts

The project is framed within applied research with the aim of developing a tool that can be integrated into digital platforms to moderate content and analyse discourse. The research is carried out in collaboration with academic institutions and could be adopted by technology companies and social media platforms.

The project has several impacts on different aspects. Socially, it supports content moderation and the study of social movements. Ethically and legally, it emphasises algorithmic accountability and user privacy. Economically, it supports automation and continuous use of the model. Environmentally, it considers energy consumption and emissions.

Aspect	**Data Collection**	**Design and Development**	**Testing**	**Implementation**
Ethical	User Privacy	Algorithmic Accountability	Data Annotation Bias	Impartiality in Evaluation
Economic	Financial Feasibility	Research-Development Costs	Improved Productivity	Project Profitability
Social	Equitable Access to Technology	Impact on Vulnerable Groups	Inclusion / Diversity	Impact digital interactions
Environmental	Resource Efficiency	Energy Efficiency	Environmental Emissions	Environmental Emissions

Table 8.1: Impacts throughout the project life cycle.

8.2 Detail of Impacts

This section details the main impacts of the project, describing each one and indicating the possible groups or sectors affected, the relevant regulations and ethical codes, the assessment possibilities and the associated economic impact.

The first relevant impact is on users' privacy, which arises mainly during the data collection phase. The collection of personal data to train models can lead to a violation of privacy, which particularly affects users of social networks. Regulations such as the GDPR (General Data Protection Regulation) are crucial in this respect; assessment options include the measurement of data breaches, which may involve costs related to the implementation of data protection measures and possible fines for non-compliance.

Another important impact is algorithmic liability, which manifests itself mainly during the design and testing phase of the system. The use of AI models for content moderation poses risks of bias and error, affecting users of digital platforms, AI developers and technology companies using the model. Regulations such as the EU's Artificial Intelligence Regulation and ISO/IEC standards on AI provide guidance in this regard. However, mitigating these concerns requires investment in the development of monitoring and auditing systems, which also has economic implications. Energy consumption and emissions are also a critical issue, especially during the training and testing phases of models. The high energy consumption of AI models, such as transformers, contributes significantly to the carbon footprint, with implications for society as a whole. Regulations such as the Zero Net Emissions Industry Act (EU) impose restrictions in this regard. Measuring energy consumption and carbon footprint during model training is one way of assessing this.

Finally, the impact on vulnerable groups is a major concern, especially during the use phase of the system. AI-based content moderation may disproportionately affect certain groups if the correct and regulated data is not included, potentially increasing pre-existing biases and affecting vulnerable communities. In this regard, system fairness assessment, model accountability, disparate impact analysis and auditing are essential to address these concerns.

8.3 Sustainable Development Goals

This project contributes to several Sustainable Development Goals (SDGs) set by the United Nations Agenda 2030:

- **SDG 3. Health and Well-being**: contributes to improving the tone and quality of online interactions, which in turn promotes the mental well-being of users.
- **SDG 4. Quality Education**: promotes awareness and dissemination of ethical and moral values, which contributes to quality education.
- **SDG 5. Gender Equality** : identifies and moderates possible discriminatory and sexist discourse, promoting gender equality in digital environments.
- **SDG 11. Sustainable Cities and Communities**: contributes to the creation of safe and more inclusive digital communities, improving the quality of public debate.
- **SDG 12. Responsible Consumption and Production**: encourages responsible consumption of digital content by regulating participation in an ethical and respectful

manner. It also improves the efficiency of content moderation, reducing the need for intensive human intervention and optimising the use of technological resources.

- **SDG 16. Peace, Justice and Strong Institutions**: helps prevent arguments by identifying and mitigating hate speech and violence online. It also promotes transparency and accountability by integrating technologies adapted to moral values.

8.4 Emission Impact Detailed

The increasingly straightforward access to pre-trained models, facilitated by platforms like Hugging Face, has democratized the use of artificial intelligence, allowing a wide variety of users to benefit from these resources without the need to train models from scratch. However, this accessibility also entails an increase in environmental impact, as the widespread use of these models implies a greater demand for computational resources and, consequently, a higher emission of greenhouse gases [19].

The environmental impact of transformer models, such as BERT and GPTs, is significant due to the substantial consumption of computational resources throughout their lifecycle, from development to deployment. The initial development phase involves designing the model architecture, conducting numerous experiments to test designs on a smaller scale, and optimizing hyperparameters. Each of these activities entails running multiple training cycles, which consumes a significant amount of electricity. The energy required for these computational tasks directly translates into greenhouse gas emissions, especially if the electricity is sourced from non-renewable sources. According to the study [86], training a single AI model like BERT can emit as much CO2 as five cars over their entire lifespan.

Additionally, the infrastructure supporting training and deployment, such as cooling systems and power supply in data centers, significantly contributes to the overall carbon footprint. According to a report by OpenAI [3], the computational power used for training large AI models has been doubling approximately every 3.4 months since 2012, leading to increased energy demands.

To mitigate these environmental impacts, it is necessary to implement strategies that optimize the development and usage of these models. This includes techniques like Bayesian optimization to reduce the number of training runs required, the use of energy-efficient hardware such as GPUs and TPUs, and selecting regions where data centers operate on renewable energy. Furthermore, it is crucial to promote transparency and disclosure of information regarding emissions associated with these models, so that users can make in-

formed and responsible decisions about their usage.

The mean emissions of the various classification tasks implemented in the study are detailed in Table 8.2, including fine-tuning and inference on the datasets including the 4 different tasks explained in 5.5. This was done using the CodeCarbon tool, which tracks several different metrics for computational costs.

	Duration (min)	**Emission (kg)**	**Emission rate (kg/s)**	**RAM power (W)**	**CPU energy (kWh)**	**GPU energy (kWh)**	**RAM energy (kWh)**	**Energy consumed (kWh)**
Fine Tuning	56.91	1.08e-03	1.90e-05	47.08	6.72e-05	3.54e-03	7.44e-04	4.95e-03
Inference	22.68	6.00e-06	1.40e-05	47.08	4.00e-03	2.00e-05	5.00e-06	2.90e-05

Table 8.2: Average Emissions and Energy Consumption for Fine-Tuning and Inference.

As shown in the table, fine-tuning the models generates significantly more emissions and energy consumption than inference tasks due to the multiple training cycles involved. The emissions generated by the use of these models reflect a significant environmental impact, especially as much of the electricity used in these processes comes from non-renewable sources. It is evident that fine-tuning and inference tasks, while critical to the adaptation and application of these models to specific contexts, contribute significantly to the carbon footprint. In particular, fine-tuning, which is essential for adapting a pre-trained model to a specific dataset, consumes significantly more energy than inference tasks.

8.5 Conclusion

In conclusion, this project has the potential to have a significant impact in a number of ways, both positive and negative. While there is the potential for negative environmental impacts, such as increased energy consumption and emissions associated with the training and implementation of Artificial Intelligence models, there are also positive aspects in social and ethical terms if constraints and potential problems are properly addressed.

From a social point of view, the project can contribute to improving the quality of online public debate, encourage more respectful and understanding communication in social networks and other digital environments, and help identify and mitigate hate speech and divisive discourse. It can also have a positive impact on online inclusion and diversity by

promoting a more equitable and safer digital environment.

On the ethical side, it is crucial to address concerns related to user privacy, algorithmic accountability and bias in the detection of moral grounds. Taking steps to ensure transparency and accountability in system development and use can help mitigate these risks and promote ethical use of technology. If the challenges are properly managed and measures are taken to mitigate negative impacts, this project has the potential to bring about positive change.

CHAPTER 9

Economic budget

This section details an adequate budget to bring about the project. Assuming a 5-year amortization period, the used resources are:

Personal Computer (Software included): This includes a high-performance personal computer necessary for the development and execution of the system. It includes essential peripherals such as a monitor, mouse, and keyboard. The total purchase price is around €1,500.00. Over the 7-month project duration, its depreciation cost is approximately €175.00.

GPU (Graphics Processing Unit): To handle the computationally intensive tasks involved in training and deploying Transformer models, a powerful GPU is required. The selected GPU costs €3,500.00, and the depreciation cost over the 7-month period is €408.33.

A summary of depreciation is shown in Table 9.1, and the total cost is indicated in Table 9.2.

Concept	Purchase Price	Use (Months)	Amortization (Years)	TOTAL
Personal Computer (Software included)	€1,500.00	7	5	€175.00
GPU	€3,500.00	7	5	€408.33

Table 9.1: Depreciation of Material Resources.

Concept	Hours	Price/hour	TOTAL
Cost of Labor(direct cost)	360	€12	€4,320
Cost of Material Resources (direct cost)			
Personal Computer (Software included)	-	-	€175.00
GPU	-	-	€408.33
TOTAL	-	-	€583.33
General Expenses (indirect costs)	-	-	€735.50
Industrial Profit	-	-	€338.33
SUBTOTAL BUDGET	-	-	€6,377.16
Applicable vat	-	-	€1,339.20
TOTAL BUDGET	-	-	€7,716.37

Table 9.2: Project Budget.

Bibliography

[1] https://ashutoshtripathi.com/2021/07/02/what-is-the-main-difference-between-rnn- [Accessed 29-05-2024].

[2] https://www.qrca.org/blogpost/1488356/Qual-Power?tag=corporate+morality. [Accessed 30-05-2024].

[3] https://openai.com/index/ai-and-compute/. [Accessed 02-06-2024].

[4] https://lilianweng.github.io/posts/2018-06-24-attention/transformer.png. [Accessed 29-05-2024].

[5] codecarbon — pypi.org. https://pypi.org/project/codecarbon/2.1.4/. [Accessed 02-06-2024].

[6] Hugging Face – The AI community building the future. — huggingface.co. https://huggingface.co/. [Accessed 29-05-2024].

[7] Matplotlib &x2014; Visualization with Python — matplotlib.org. https://matplotlib.org/. [Accessed 29-05-2024].

[8] pandas - Python Data Analysis Library — pandas.pydata.org. https://pandas.pydata.org/. [Accessed 29-05-2024].

[9] scikit-learn: machine learning in Python &x2014; scikit-learn 1.5.0 documentation — scikit-learn.org. https://scikit-learn.org/stable/. [Accessed 29-05-2024].

[10] seaborn: statistical data visualization &x2014; seaborn 0.13.2 documentation — seaborn.pydata.org. https://seaborn.pydata.org/. [Accessed 29-05-2024].

[11] Welcome to Python.org — python.org. https://www.python.org/. [Accessed 29-05-2024].

[12] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184, 2020.

[13] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184, March 2020.

[14] Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques, 2018.

[15] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[16] Lina Bentahila, Roger Fontaine, and Valérie Pennequin. Universality and cultural diversity in moral reasoning and judgment. *Frontiers in Psychology*, 12:764360, 2021.

[17] Axel Bruns and Jean Burgess. Crisis communication in natural disasters: The queensland floods and christchurch earthquakes. *Twitter and society [Digital formations, Volume 89]*, pages 373–384, 2014.

[18] Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovì. Detection of morality in tweets based on the moral foundation theory. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 1–13. Springer, 2022.

[19] Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. Exploring the carbon footprint of hugging face's ml models: A repository mining study. pages 1–12, 10 2023.

[20] Qingqing Chen and Andrew Crooks. Analyzing the vaccination debate in social media data pre- and post-covid-19 pandemic. *International Journal of Applied Earth Observation and Geoinformation*, 110:102783, 2022.

[21] Zheng Chen and Yunchen Zhang. Better few-shot text classification with pre-trained language model. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30*, pages 537–548. Springer, 2021.

[22] Molly J Crockett. Moral outrage in the digital age. *Nature human behaviour*, 1(11):769–771, 2017.

[23] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[24] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[26] Theodoros Evgeniou and Massimiliano Pontil. Support vector machines: Theory and applications. volume 2049, pages 249–257, 09 2001.

[27] Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, 2016.

[28] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information sciences*, 180(10):2044–2064, 2010.

[29] Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*, 2016.

[30] Homero Gil de Zúñiga, Nakwon Jung, and Sebastián Valenzuela. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of computer-mediated communication*, 17(3):319–336, 2012.

[31] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE, 2011.

[32] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

[33] Jonathan Haidt. Morality. *Perspectives on psychological science*, 3(1):65–72, 2008.

[34] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

[35] Jonathan Haidt and Craig Joseph. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66, 2004.

[36] Steffen Herbold. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173, 2020.

[37] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[38] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

[39] Joseph Hoover, Kate M Johnson-Grey, Morteza Dehghani, and Jesse Graham. Moral values coding guide, Jul 2017.

[40] Xiaolei Huang, Alexandra Wormley, and Adam Cohen. Learning to adapt domain shifts of moral values via instance weighting. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 121–131, 2022.

[41] Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. 2012.

[42] Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*, 2022.

[43] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[44] Andreas Jungherr. Analyzing political communication with digital trace data. *Cham, Switzerland: Springer*, 2015.

[45] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[46] Kyriaki Kalimeri, Mariano G Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, 92:428–445, 2019.

[47] Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. Human values and attitudes towards vaccination in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 248–254, 2019.

[48] Rishemjit Kaur and Kazutoshi Sasahara. Quantifying moral foundations from various topics on twitter conversations. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2505–2512. IEEE, 2016.

[49] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. Moral concerns are differentially observable in language. *Cognition*, 212:104696, 2021.

[50] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30, 2022.

[51] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale, Jul 2018.

[52] Anna Koufakou and Jason Scott. Lexicon-enhancement of embedding-based approaches towards the detection of abusive language. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 150–157, 2020.

[53] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. Frameaxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644, 2021.

[54] Alex Gwo Jen Lan and Ivandré Paraboni. Text-and author-dependent moral foundations classification. *New Review of Hypermedia and Multimedia*, 28(1-2):18–38, 2022.

[55] Ci-Jyun Liang, Thai-Hoa Le, Youngjib Ham, Bharadwaj RK Mantha, Marvin H Cheng, and Jacob J Lin. Ethics of artificial intelligence and robotics in the architecture, engineering, and construction industry. *Automation in Construction*, 162:105369, 2024.

[56] Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, pages 552–559. IEEE, 2018.

[57] Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn M Jonker, Kyriaki Kalimeri, and Pradeep K Murukannaiah. What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric. In *PROCEEDINGS OF THE 61ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2023): LONG PAPERS, VOL 1*, pages 14113–14132, 2023.

[58] Enrico Liscio, Alin E Dondera, Andrei Geadau, Catholijn M Jonker, and Pradeep K Murukannaiah. Cross-domain classification of moral values. In *2022 Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745. Association for Computational Linguistics (ACL), 2022.

[59] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. What values should an agent align with? an empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1):23, 2022.

[60] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. Axies: Identifying and evaluating context-specific values. In *AAMAS*, pages 799–808, 2021.

[61] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[62] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023.

[63] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[64] Gilles Louppe. Understanding random forests: From theory to practice, 2015.

[65] Mia Lövheim, André Jansson, Susanna Paasonen, and Johanna Sumiala. Social media: implications for everyday life, politics and human agency. 2013.

[66] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[67] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[68] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[69] Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 206–219. Springer, 2020.

[70] Ece Çiğdem Mutlu, Toktam Oghaz, Ege Tütüncüler, and Ivan Garibay. Do bots have moral judgement? the difference between bots and humans in moral rhetoric. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 222–226. IEEE, 2020.

[71] Anny D Alvarez Nogales and Oscar Araque. Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 67–77, 2024.

[72] Tunboson Oyewale Oladoyinbo, Samuel Oladiipo Olabanji, Oluwaseun Oladeji Olaniyi, Olubukola Omolara Adebiyi, Olalekan J Okunleye, and Adegbenga Ismaila Alao. Exploring the challenges of artificial intelligence in data integrity and its influence on social dynamics. *Asian Journal of Advanced Research and Reports*, 18(2):1–23, 2024.

[73] Jeongwoo Park, Enrico Liscio, and Pradeep K Murukannaiah. Morality is non-binary: Building a pluralist moral sentence embedding space using contrastive learning. *arXiv preprint arXiv:2401.17228*, 2024.

[74] Marco Pennacchiotti and Ana-Maria Popescu. Detecting controversies in twitter: a first study. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*, pages 31–32, 2010.

[75] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[76] Vjosa Preniqi, Iacopo Ghinassi, Kyriaki Kalimeri, and Charalampos Saitis. Moralbert: Detecting moral values in social discourse, 2024.

[77] Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 177–188, 2021.

[78] Rezvaneh Rezapour, Saumil H Shah, and Jana 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, 2019.

[79] Shamik Roy and Dan Goldwasser. Analysis of nuanced stances and sentiment towards entities of us politicians through the lens of moral foundation theory. In *Proceedings of the ninth international workshop on natural language processing for social media*, pages 1–13, 2021.

[80] Eyal Sagi and Morteza Dehghani. Moral rhetoric in twitter: A case study of the us federal shutdown of 2013. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

[81] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[82] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*, 2019.

[83] James Shaw, Joseph Ali, Caesar A Atuire, Phaik Yeong Cheah, Armando Guio Español, Judy Wawira Gichoya, Adrienne Hunt, Daudi Jjingo, Katherine Littler, Daniela Paolotti, et al. Research ethics and artificial intelligence for global health: perspectives from the global forum on bioethics in research. *BMC Medical Ethics*, 25(1):46, 2024.

[84] Rui Song, Zelong Liu, Xingbing Chen, Haining An, Zhiqi Zhang, Xiaoguang Wang, and Hao Xu. Label prompt for multi-label text classification. *Applied Intelligence*, 53(8):8761–8775, 2023.

[85] Jacopo Staiano and Marco Guerini. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*, 2014.

[86] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019.

[87] Khairul Syafuddin. Ethics of artificial intelligence: Dialectics of artificial intelligence policy for humanity. *The Eastasouth Journal of Information System and Computer Science*, 1(03):147–154, 2024.

[88] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[89] Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*, 2022.

[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[91] Dragos Vecerdea. Moral embeddings: A closer look at their performance, generalizability and transferability. 2021.

[92] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835, 2018.

[93] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. *Twitter and society*. Peter Lang New York, 2013.

[94] Lorraine Whitmarsh. Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global environmental change*, 21(2):690–700, 2011.

[95] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[96] Jing Yi Xie, Graeme Hirst, and Yang Xu. Contextualized moral inference. *arXiv preprint arXiv:2008.10762*, 2020.

[97] Wang Yaqing, Yao Quanming, T Kwok James, and M Ni Lionel. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys*, 53(3):1–34, 2020.

[98] Wei Zhu and Daniel Cheung. Lex-bert: Enhancing bert based ner with lexicons. *arXiv preprint arXiv:2101.00396*, 2021.