Universiteit Leiden

Master Computer Science

Domain-adversarial moral foundation prediction for long EU legal documents

Name:	Jakob Lindscheid
Student ID:	s3942716
Date:	July 8, 2025
Specialisation:	Data Science
1st supervisor:	Suzan Verberne
2nd supervisor:	Armin Cuyvers

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

ABSTRACT

This thesis presents a large-scale, context-aware approach for predicting moral foundations in European Union (EU) law documents using pretrained Transformer-based language models. Prior research has found that a lack of moral diversity in the EU's laws and communications may impact its perceived legitimacy, and therefore effectiveness. Existing analyses have been limited in scale and only employed traditional dictionary-based methods. We address this gap by finetuning a pretrained Transformer encoder for moral foundation prediction. To transfer moral foundation knowledge from labeled social media and news datasets to the unlabeled EU documents, we apply domain-adversarial training, which encourages the learning of domain-invariant features. For evaluation purposes, we introduce a manually annotated dataset of 111 EU law documents labeled for moral foundations.

In developing our model, we address two central challenges: 1) the prevalent class imbalance between positive and negative samples, and 2) the considerable length of EU documents. For the first, we demonstrate that class weights, focal loss, and training with an increased batch size substantially improve model performance, especially when evaluating on cross-domain data. For the second, we propose two solutions: aggregating the predictions on shorter chunks of the documents and using Longformer for full-context modeling. Our results show that label aggregation achieves better performance and outperforms dictionary- and frequency-based baselines on long documents.

Applying our trained models, we confirm prior hypotheses that EU law overrepresents individualizing moral foundations (Care, Fairness) compared to binding ones (Loyalty, Authority, Purity). Our work contributes a novel dataset, a pipeline for moral foundation classification of EU documents, and an analysis of techniques for handling document length and class imbalance. This advances the field of moral foundation prediction and provides tools for future research at the intersection of natural language processing (NLP), political science, and psychology.

CONTENTS

1 INTRODUCTION

The perceived legitimacy of supranational organizations like the European Union (EU) is a key influence on their stability and effectiveness. Specifically, the resources (mandates, participation, and financial resources), the policy output, and the rule compliance within the institutions depend on the public's acceptance of their authority (Sommerer and Agné, 2018).

According to the Eurobarometer (Spring 2025)[1] only 52% of EU citizens tend to trust the EU. Furthermore, eurosceptic parties have established themselves as a permanent part of European and national politics in the last decade (Treib, 2021). At the same time, the need for public legitimacy grows with the required effectiveness of the EU in responding to recent transnational problems such as the COVID-19 pandemic, climate change, intensified immigration issues, and a changing geopolitical landscape with a possibly decreased reliance on the United States.

Several studies have shown that the perceived legitimacy of an authority is linked to the belief in the moral rightness of its actions and laws (Bottoms and Tankebe, 2012; Jackson, Bradford, et al., 2012; Jackson, Hough, et al., 2015; Tyler and Jackson, 2013). Based on this, Grosfeld et al. (2024) suggest that the EU's public legitimacy problem is partly rooted in a lack of moral diversity in its communications and laws. To analyze the morality of the EU, they employ Moral Foundations Theory (MFT) (Graham, Haidt, Koleva, et al., 2013; Haidt and Joseph, 2004). This theory suggests that human morality is rooted in several innate psychological systems, which have developed in humans because of evolutionary advantages, mostly in creating a functioning society. The original five foundations are Care, Fairness, Loyalty, Authority and Purity. According to MFT, these five cognitive moral learning modules are present at birth, but individual experiences (upbringing, education, culture, ...) determine which moral foundations become more important to people.

To support their hypothesis of the EU's lacking moral diversity, Grosfeld et al. (2024) conduct several studies and explore moral reframing of EU law. One of these studies is a relatively small-scale text analysis of State of the Union speeches using a dictionary-based word frequency analysis. The results of this study show that there is a significantly higher probability that moral words in the speeches were related to individualizing moral foundations (care & fairness) than being related to binding moral foundations (loyalty, authority & purity).

1 https://europa.eu/eurobarometer/surveys/detail/3372

This study is fundamentally relevant and its results substantial. However, it leaves potential for a much larger-scale analysis of EU law texts. State of the Union speeches are limited to the position of the Commission President and the amount of analyzable text is relatively small. Furthermore, dictionary-based methods such as the one used in the study are relatively simple and cannot take the context a word is used in into account. A single word can also have multiple, completely different meanings, and especially EU documents tend to use words outside their everyday context. For example, regulations that discuss the trade of "oilseed rape" may be detected as extremely moral, because the term "rape" is used a lot even though the document has no moral content at all.

Furthermore, Moral Foundation Dictionaries are commonly developed by groups of experts (Graham, Haidt, and Nosek, 2009) or generated based on annotations of moral documents such as news texts (Hopp et al., 2021). These approaches cannot take every type of document into account and therefore do not necessarily generalize well to EU law documents.

For a larger-scale analysis, we require a context-aware approach that is specifically developed to determine the moral foundations expressed by EU law documents. Modern natural language processing (NLP) techniques, especially pretrained Transformer-based language models (Devlin et al., 2019; Vaswani et al., 2017) can address these technical challenges. In this thesis, we finetune such a model for moral foundation prediction in EU law documents by transferring the knowledge learned from labeled social media and news texts to unlabeled EU documents.

Applying Transformer models to the task of moral foundation prediction is not a new idea (Kobbe et al., 2020; Roy and Goldwasser, 2021; Trager et al., 2022; Zangari et al., 2025). These previous works are limited to few available labeled datasets that contain textual content from social media (Beiró et al., 2023; Hoover et al., 2020; Johnson and Goldwasser, 2018; Trager et al., 2022) or news texts (Hopp et al., 2021; Weber et al., 2021). Models that are trained on a specific data domain (e.g. social media or news) may not work well when applied to a substantially different domain such as EU law texts. Liscio et al. (2022) observe this effect even when transferring knowledge between various Twitter corpora that cover different topics.

To the best of our knowledge, there is currently no dataset of EU documents that is labeled for moral foundations. Therefore, we adopt the unlabeled domain transfer approach developed by Guo et al. (2023) and Preniqi et al. (2024) to transfer the knowledge learned from a labeled source domain (social media, news) to an unlabeled target domain (EU documents). Additionally, we create a small labeled dataset of 111 EU law documents that we use to evaluate the trained document.

Guo et al. (2023) introduce the idea of weighting the loss function of a Transformer to balance the different classes and between positive and negative data samples. We consider the class imbalance as a fundamental challenge in moral foundation prediction, especially in the cross-domain setting. To address this challenge, we suggest training

with a larger batch size and exchanging the typically used cross-entropy loss function with focal loss (T.-Y. Lin et al., 2017), which is an extension of the former that was developed to address the extreme class imbalance observed in dense object detection.

Finally, we address the challenge that EU documents tend to be extremely long. In Section 3.2, we see that only 11.3% of the considered EU documents fit into the 512 token context window of commonly used Transformer encoders like BERT (Devlin et al., 2019) and RoBERTa (Y. Liu et al., 2019). To the best of our knowledge, no previous work has addressed this limitation in moral foundation prediction. We propose and compare two solutions for determining the moral foundations expressed in long documents. For standard BERT models, we split the documents into smaller chunks that fit into the respective context window and then aggregate the model outputs to retrieve predictions for the whole document. As a second approach, we use Longformer (Beltagy et al., 2020), which uses local windowed attention to process longer sequences efficiently.

In the following, we summarize our four main contributions:

1. We conduct a large-scale study of the moral foundations expressed by EU law documents. For that, we train a Transformer model using a domain transfer approach to transfer the knowledge learned on labeled datasets to the unlabeled EU documents.

2. We provide a small dataset of 111 EU law documents labeled with the expressed moral foundations.

3. We evaluate three solutions to the inherent class imbalance in moral foundation prediction.

4. To the best of our knowledge, we are the first to address the task of moral foundation prediction in long documents (> 512 tokens). We propose two fundamentally different approaches for long documents and discuss the implications of their results for moral foundations theory in general.

2 BACKGROUND

2.1 MORAL FOUNDATIONS THEORY

Moral Foundations Theory (MFT) is a social psychological model of human morality primarily developed by Haidt and Joseph (2004). It unifies several previous theories (Fiske, 1991; Schwartz and Bilsky, 1990; Shweder, 1987) in an effort to explain similarities in morality across populations despite substantial differences between cultures. MFT suggests that human morality is rooted in several innate psychological systems, which have developed in humans because of evolutionary advantages, mostly in creating a functioning society. The differences across cultures are explained by cultures building their virtues, institutions, and religions on differing distributions of these foundations. On an individual level, MFT argues that the moral foundations are present in the brain at birth but then develop based on the individual environment such as upbringing, education and own experiences.

There are four main claims that summarize MFT as outlined by Graham, Haidt, Koleva, et al. (2013):

1. Nativism: There is a "first draft" of the moral mind.
2. Cultural learning: The first draft gets edited during development within a particular culture.
3. Intuitionism: Intuitions come first, strategic reasoning second.
4. Pluralism: There are many recurrent social challenges, so there are many moral foundations.

Nativism and cultural learning are essentially explained above. Intuitionism refers to the Social Intuitionist Model (Haidt, 2001), which proposed that most moral evaluations happen rapidly and automatically (i.e. System 1 thinking following Kahneman (2011)), while deliberative, logical thinking (System 2) is more often initiated to explain, defend, or justify intuitive moral reactions. Finally, pluralism refers to the development of moral foundations through evolution as solutions to social challenges.

At its core, MFT makes no claim about specific foundations or even the number of foundations. However, five foundations were published along with the theory that had the best evidence according to the authors (Graham, Haidt, Koleva, et al., 2013). Each foundation consists of virtues and vices, which are each summarized in a single word

that best represents the support (virtue) or violation (vice) of the respective foundation. The original five moral foundations are the following:

- **The Care/harm foundation** is the connection between perceptions of suffering and the motivation to care, nurture and protect. The related evolutionary advantage is in caring for ones own children, but in an established society the effect goes clearly far beyond that.
- **The Fairness/cheating foundation** is the sensitivity to evidence of cheating and cooperation usually in opportunities for mutually beneficial cooperation. Having an intuitive understanding of fairness is an advantage when trying to benefit from cooperation without being exploited.
- **The Loyalty/betrayal foundation** refers to the ability to form and maintain coalitions, especially in identifying team-players and traitors. Effective tribalism has been an (evolutionary) advantage when faced with challenges and attacks from rival groups. A modern example for this foundation are sports fandom and brand loyalty.
- **The Authority/subversion foundation** refers to the ability to perceive and react appropriately to a hierarchical order. This is two-fold, as one can be either the superior or subordinate in each situation. This foundation faces the social challenge of forging beneficial relationships in hierarchies, which have been observed in human tribes and early civilizations.
- **The Sanctity/degradation foundation** in essence describes the sense of disgust. It includes the "behavioral immune system", which increases wariness towards symbolic objects and threats. This foundation most likely evolved to avoid pathogens and parasites in food or waste.

While these are the original five foundations, the authors of MFT (mainly Haidt and Joseph (2004) and Graham, Haidt, Koleva, et al. (2013)) stress several times that they do not believe that these are the only moral foundations. For example, Haidt (2012) considers adding Liberty/Oppression as a sixth foundation to better capture the moral concerns of libertarians. Graham, Haidt, Koleva, et al. (2013) also mention Efficiency/waste, Ownership/theft and Honesty/deception as good candidates for new moral foundations and outline four criteria that foundations should fulfill. Recently, Atari et al. (2023) suggested splitting the fairness foundation into equality and proportionality. Since this is an ongoing debate in the field of psychology and most follow-up work to MFT focuses on the original five foundations, this thesis will also be limited to those. Specifically, we determine whether a text supports or violates each of these foundations, i.e. we distinguish between virtues and vices.

Graham, Haidt, Koleva, et al. (2013) argue for the pragmatic validity of MFT by presenting findings enabled by MFT, covering political ideology, relations between moral

foundations and other psychological constructs, cross-cultural differences, intergroup relations, and implicit processes in moral cognition. Four main methods have been used to measure the MFT: 1) self-report surveys (Graham and Haidt, 2012; Graham, Nosek, et al., 2011), 2) implicit measures (Graham, 2010), 3) psychophysiological and neuroscience methods (Cannon et al., 2011; Graham, 2010), and 4) text analysis (Graham, Haidt, and Nosek, 2009; see Section 2.2). For this thesis, the most relevant aspects are text analysis and the connection between moral foundations and political orientation.

While MFT was not developed for political psychology, Haidt and Graham (2007) suggest that it may be used to explain differences in political orientation. They make the prediction that liberals more heavily rely on the Care and Fairness foundations ("individualizing" foundations) than conservatives, who would rather rely on the Loyalty, Authority and Sanctity foundations ("binding" foundations). Several subsequent studies find support for this prediction. Graham, Haidt, and Nosek (2009) consistently observe this pattern across four different methods (3 self-report questionnaires and 1 text analysis of church sermons), Graham, Nosek, et al. (2011) confirm this finding using responses to the Moral Foundations Questionnaire (MFQ)[1] from 11 different world regions and McAdams et al. (2008) find "strong support" for MFT in life narrative interviews with highly religious and political-engaged adults.

2.2 RELATED WORK

TRADITIONAL TEXT ANALYSIS IN MFT Text analysis has been a common method in researching MFT since its original conception. As mentioned in Section 2.1, Graham, Haidt, and Nosek (2009) analyzed church sermons to confirm the connection between moral foundations and political orientation. For this text analysis, they developed the first Moral Foundations Dictionary (MFD) consisting of 324 words and word stems that are assigned to one or more moral foundation if they express the support or violation (virtue/vice) of that foundation. The MFD was developed by a group of experts by generating words related to the base foundation words (e.g. care, harm, etc.). Several works use this dictionary to analyze the morality of texts (Dehghani et al., 2014; Lewis, 2019; Takikawa and Sakamoto, 2017).

THE EXTENDED MORAL FOUNDATIONS DICTIONARY The MFD has been expanded by various works with different techniques (Araque et al., 2020; Frimer et al., 2019). Recently, the "extended Moral Foundations Dictionary" (eMFD; Hopp et al., 2021) has been introduced that addresses several limitations of previous works. Over 800 annotators were provided with news articles that were published between November 2016 and January 2017 by different US news outlets. The annotators were then asked

1 https://www.yourmorals.org/

to highlight parts of the text that reflect their assigned moral foundation. Hopp et al. (2021) use these annotations to compute the probabilities that a certain term is annotated for each moral foundation. This results in continuous vectors for more than 3,000 terms.

In addition to the eMFD itself, Hopp et al. (2021) provide a document scoring method, which computes a document score by averaging the probabilities of all words in a document. While this method could be used to score EU law documents, dictionary-based methods have several limitations that our approach addresses. For our work, the most relevant limitation of dictionaries is their inability to take context into account. Since dictionary entries are single words, every method based on these entries can only use the presence of words in a document, but their position and the context they are used in cannot be taken into account. We address this challenge by employing Transformer-based models, which are capable of capturing contextual relationships between words by considering their positions and dependencies within a sequence (Vaswani et al., 2017).

ANALYZING LEGISLATURE Since Graham, Haidt, and Nosek (2009) identified the connection of MFT with political orientation, research also started using MFT to analyze the legislative process. However, instead of examining law texts directly, previous works rather use related documents such as legislative speeches (Mucciaroni, 2011), letters to the editor (Burlone and Richmond, 2018), or voter guides (Wendell and Tatalovich, 2021). Furthermore, research in this area has mostly focused on the United States. Notable exceptions are Harper and Hogue (2019), who analyze campaign news and speeches surrounding the Brexit referendum, and Grosfeld et al. (2024), who apply the eMFD to State of the Union speeches to explore the presence/lack of moral diversity in the EU. To the best of our knowledge, all previous analyses have either inspected documents manually or used dictionary-based methods like the ones described above. Therefore, we contribute by analyzing law texts directly, further expanding the field beyond the US, and employing context-aware Transformer models instead of traditional methods.

TRANSFORMERS FOR MORAL FOUNDATION PREDICTION The first work to apply pretrained Transformer models to moral foundation prediction was done by Kobbe et al. (2020), who compare their lexicon-based approaches to a BERT model (Devlin et al., 2019) finetuned for multi-label classification of the five original moral foundations. They find that the BERT model outperforms all other tested approaches on the ArgQuality Corpus of Wachsmuth et al. (2017) and the Moral Foundations Twitter Corpus (MFTC; Hoover et al., 2020; see Section 3.1.2) . Also using the MFTC, Bulla et al. (2022) confirm the outperformance of the same BERT-based method compared to an LSTM-based approach for 10 dimensional moral foundation prediction, i.e. with the added aspect of virtues and vices (care/harm, fairness/cheating, etc.).

While providing baselines for their Moral Foundations Reddit Corpus (MFRC; see Section 3.1.3), Trager et al. (2022) introduce the idea of training one binary classifier for each label to create an ensemble of single-label classifiers. They find that this approach outperforms the previously used multi-label method. This finding is later confirmed by Nguyen et al. (2024) and Preniqi et al. (2024).

DOMAIN ADAPTATION Both Trager et al. (2022) and Liscio et al. (2022) examine the domain transfer challenge in moral foundation prediction by finetuning a BERT model on data from a source domain and evaluating it on data from a target domain. Trager et al. (2022) use Tweets and Reddit posts as two different domains and employ the single-label ensemble approach as described above. Liscio et al. (2022) apply the earlier multi-label method and consider the seven corpora within the MFTC as different domains. While both works find that cross-domain classification is fundamentally possible, they also observe substantial performance impacts compared to in-domain classification. Depending on the setting and label, the average F1 score drops by approximately 10% for Liscio et al. (2022) and up to 52% for Trager et al. (2022).

To address this challenge Guo et al. (2023) propose a Domain Adapting Moral Foundation inference model (DAMF), which uses domain adversarial training (Ganin et al., 2016) to learn domain-invariant features. A domain adversarial neural network consists of a feature extractor that feeds into a label classifier and an adversary domain classifier, which are each trained to identify the labels or domain of a data sample, respectively. The feature extractor is trained to optimize label classification while compromising the domain classification, which results in domain-invariant features for label classification. DAMF uses this setup with BERT as a feature extractor, a label classification head for multi-label moral foundation prediction and a domain classification head to distinguish source domain from target domain. Furthermore, Guo et al. (2023) suggest adding three more components: a domain-invariant transformation, a reconstruction module and a weighted loss function that accounts for class imbalance. We describe this architecture in more detail in Section 4.4. DAMF achieves substantially better results in the domain transfer setting compared to a standard BERT model that is finetuned on the source domain. Additionally, Guo et al. (2023) observe that a large part of the improvement is achieved by accounting for class imbalance through the weighted loss function. DAMF improves the F1 score by 31.6% on average over the standard BERT method, but without the weighted loss function the improvement is only 9.8% on average.

Preniqi et al. (2024) extend the work of Guo et al. (2023) by applying the DAMF architecture to the idea of using an ensemble of single-label models as we described above. Since we adopt the DAMF architecture and compare the multi-label and single-label ensemble approaches, the works of Guo et al. (2023) and Preniqi et al. (2024) are the most relevant to our research. Both also distinguish between virtues and vices, i.e. predict 10 labels. We apply these approaches to the gap of moral foundation prediction

in EU law documents as described above. Furthermore, we contribute to the field of moral foundation prediction in two more ways. First, we pick up the idea of Guo et al. (2023) to account for class imbalance in moral foundations, which we consider as a fundamental challenge. In addition to class weights, we suggest employing focal loss (T.-Y. Lin et al., 2017) instead of the standard cross-entropy loss and training with a higher batch size to counteract the invalid sampling problem. Second, we address moral foundation prediction in long documents (> 512 tokens) by proposing two fundamentally different approaches.

3 DATA

In this section, we introduce the different datasets used in this work. First, we describe existing labeled datasets: the Moral Foundations News Corpus (MFNC; Section 3.1.1), which we generate from the annotation effort by Hopp et al. (2021), the Moral Foundations Twitter Corpus (MFTC; Section 3.1.2), and the Moral Foundations Reddit Corpus (MFRC; Section 3.1.3). Then, in Section 3.2.1 we explain which EU documents we extracted from the EU's data repository to build a corpus of EU documents for our analysis. In Sections 3.2.2, 3.2.3, and 3.2.4, we describe our annotation effort, which results in 111 labeled documents. Finally, we summarize all presented datasets in Section 3.3 and conduct some exploratory analysis.

3.1 LABELED MORAL FOUNDATIONS DATA

3.1.1 Annotations from extended Moral Foundations Dictionary

To construct the eMFD (see Section 2.2), Hopp et al. (2021) asked over 800 annotators to highlight parts of news articles that reflect a certain moral foundation. We use these annotations to construct two datasets.

The annotations are supposed to generally reflect the respective moral foundation, which is independent from whether the text is supporting or violating the foundation. To compute a continuous valence score Hopp et al. (2021) used the Valence Aware Dictionary and sEntiment Reasoner (VADER; Hutto and Gilbert, 2014). Using this score, the virtue/vice aspect can be added to the five annotated classes, i.e. documents with a positive sentiment score are assigned the respective virtue, while a negative sentiment leads to the vice label. Like this, we obtain the desired 10–dimensional labels for each annotation.

Each annotation only has a single label, but some of them overlap, and we are also interested in texts that contain several sentences. So, by considering all annotations in a certain document or paragraph of a document, we can determine all moral foundations reflected by the respective text. While most of these longer texts contain annotations, not all of them do, leading to no assigned label. However, this is actually desired since a model should capture the case that a certain text does not concern any moral foundations.

In conclusion, we generate two datasets from the eMFD annotations: The 990 full documents with all corresponding annotations and the 20,427 paragraphs in these doc-

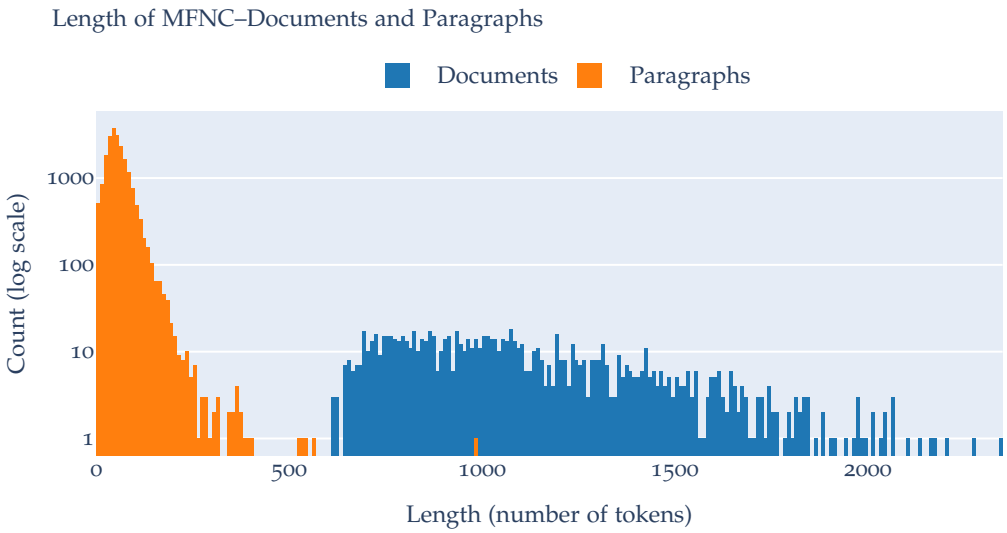

Figure 1: Length distributions of the MFNC–documents and MFNC–paragraphs datasets. Text length is measured in tokens.

uments with the annotations concerning the specific paragraph. The main differences are 1) the number of samples and 2) the length of the texts. While most of the paragraphs fit in the 512 token context window of BERT (Devlin et al., 2019), the documents do not and can therefore be considered "long documents" (Alva Principe et al., 2025). The length distributions of both datasets can be seen in Figure 1.

In our work, we will refer to the two datasets as "MFNC–paragraphs" and "MFNC–documents". The term MFNC (Moral Foundations News Corpus) was first used by Mokhberian et al. (2022) to refer to the eMFD annotations as a dataset, but our way of processing the annotations is more similar to the work of Nguyen et al. (2024), who split the documents into sentences.

3.1.2 Moral Foundations Twitter Corpus (MFTC)

The MFTC (Hoover et al., 2020) consists of over 30,000 tweets about 7 different morally relevant topics. Each tweet was annotated by at least 3 annotators, who could choose between one or multiple of the original five moral foundations (virtues, e.g. care, fairness), their violations (vices, e.g. harm, cheating) and the content being nonmoral. The MFTC therefore already fulfills all desired properties for this work.

Generally, terms of use of X (formerly Twitter) forbid publishing the text of tweets.[1] Only the corpus from Davidson et al. (2017) is available online, including the tweet texts. For the other corpora we scraped the tweets that are still available. Tweets in the MFTC that were deleted by the user or removed for other reasons can no longer be

1 https://x.com/en/tos

Corpus	Number of available tweets	Number of tweets in the original dataset
All Lives Matter	2,328 (52.6%)	4,424
Black Lives Matter	2,874 (54.7%)	5,257
Baltimore protests	2,763 (49.4%)	5,593
Hate speech & offensive language (Davidson et al., 2017)	4,873 (100%)	4,873
2016 US presidential election	3,069 (57.3%)	5,358
Hurricane Sandy	2,712 (59.1%)	4,591
#MeToo	0 (0.00%)	4,891
Total	18,619 (53.2%)	34,987

Table 1: Size of each corpus in the MFTC and how many of theses tweets we were able to scrape.

retrieved, which results in a reduced dataset size. In Table 1, we show for each corpus how many tweets are available for this work in comparison to the original amount. Since the corpora slightly overlap, we end up with 17,990 unique tweets.

We were not able to retrieve any tweets from the #MeToo corpus, which may be surprising as the percentage of retrieved tweets is relatively stable for the other corpora. However, according to Hoover et al. (2020), the #MeToo corpus was sampled by selecting 200 users that were involved in the #MeToo movement, while the selection in other corpora is usually more broad, i.e. all followers of specific politicians or all tweets containing a specific hashtag. Therefore, we assume that the number of unique users in the other corpora is much higher. We conclude that all 200 users in the #MeToo corpus must have deleted their accounts/tweets or had their tweets removed for different reasons. At such a low number of users, this is not as improbable as initially suspected. The recent work of Preniqi et al. (2024) also excludes the #MeToo corpus, most likely due to the same issue.

3.1.3 Moral Foundations Reddit Corpus (MFRC)

The MFRC (Trager et al., 2022) is a collection of 17,886 English Reddit comments from 11 different subreddits.[2] Its primary aim is to contribute a dataset from a social media platform other than Twitter. Trager et al. (2022) argue that different social media platforms differ in moral language and behavior because of different linguistic, social, and technical (e.g. character limits) environments.

The annotation process of the MFRC differs in two aspects from the MFTC. First, it does not account for the polarity of moral foundations, i.e. there is no difference

2 The number of posts and subreddits is determined based on the published dataset. In their paper, Trager et al. (2022) give different numbers: 16,123 posts from 12 subReddits.

between virtue and vice. Second, Trager et al. (2022) follow the recent revision of MFT by Atari et al. (2023) in splitting the fairness foundation into the equality and the proportionality foundations.

Since we are only interested in the original five foundations, we recombine the equality and proportionality foundations into the fairness foundation. To address the missing polarity annotations, we use the Valence Aware Dictionary and sEntiment Reasoner (VADER; Hutto and Gilbert, 2014) as it was also done by Hopp et al. (2021) in creating the eMFD. We compute the compound sentiment scores for all documents and assign the corresponding vice labels to documents with negative scores; otherwise, the virtue labels are assigned. For example, a document that receives a negative sentiment score and is labeled with "care" and "loyalty" according to the MFRC, will be labeled "harm" and "betrayal" by us. However, if the sentiment score is positive, the virtue labels will remain.

The main drawback of this approach is that it is impossible for a document to be labeled with a virtue and a vice at the same time (e.g. harm & authority), which does happen in the other datasets (9.08% of MFNC–paragraphs, 19.66% of the MFTC). The MFNC datasets do not have this issue, since VADER is applied to the individual annotations, which are combined to obtain the document/paragraph labels. Each document/paragraph can contain annotations with different sentiments according to VADER.

Since the MFTC contains explicit labels for the virtues and vices, we can validate the usage of VADER by comparing its predictions to the MFTC labels. For that, we first reduce the MFTC labels to five dimensions by grouping each virtue and vice together. Then, we expand to 10 dimensions using VADER in the same procedure as for the MFRC. The original MFTC labels are used as ground-truth to evaluate the predicted labels by VADER. The results of this validation can be seen in Table 2. From this we conclude that the chosen approach is reasonable, especially considering the lack of better alternatives. However, it should be noted that the quality of the labels is substantially degraded by using sentiment as a proxy for distinguishing virtues and vices.

Other than the reduced set of labels, the MFRC follows the MFTC in allowing multiple labels for each text and also annotating for nonmoral content.

3.2 EU DOCUMENTS

3.2.1 Document selection

To extract relevant EU documents, we use a slightly modified version of the EU Regulation Corpus Compiler developed by Seppälä (2019).[3] The EU publishes all its docu-

3 https://github.com/JakobLindscheid/eu_corpus_compiler

	Precision	Recall	F1-score	Support
Care	0.766	0.759	0.762	3,127
Harm	0.892	0.737	0.807	3,734
Fairness	0.729	0.683	0.705	2,591
Cheating	0.852	0.708	0.774	3,100
Loyalty	0.821	0.774	0.797	2,811
Betrayal	0.721	0.669	0.694	1,869
Authority	0.739	0.639	0.686	2,409
Subversion	0.685	0.593	0.636	2,096
Purity	0.733	0.803	0.767	1,307
Degradation	0.841	0.703	0.766	1,429
Macro average	0.778	0.707	0.739	24,473

Table 2: Classification metrics for the validation of using sentiment as a proxy for distinguishing virtues and vices. The 10 dimensional MFTC labels are first reduced to 5 dimensions by grouping each vice to its corresponding virtue. Then, VADER (Hutto and Gilbert, 2014) sentiment scores are used to reintroduce the distinction. The resulting labels (predictions) are compared to the original MFTC labels (ground-truth)

ments in "Cellar", the common data repository of the Publications Office of the EU.[4] Behind Cellar there is an extensive data model providing structure, metadata and links between publications. We use this structure to filter the tens of millions of available documents to those we are interested in for this work. The EU publishes every document in all official languages of the EU (currently 24). We limit this work to English documents, as the other datasets we use are also in English. While the availability of this amount of multi-lingual documents presents an interesting opportunity, we leave its exploration to future work.

The two other fields we use to filter the documents before extraction are the document type and the Eurovoc concept. In both cases, we use expert judgments by a Professor of European Law to decide which documents are relevant for analyzing moral foundations. We identify 64 different document types of which we are extracting treaties, regulations, implementing regulations, decisions, implementing decisions, international agreements, and preparatory acts.

Eurovoc[5] is a thesaurus managed by the Publications Office of the European Union. The thesaurus contains keywords that cover the activities of the EU and are organized in 21 domains and 127 subdomains. Each subdomain contains multiple "Eurovoc concepts". To filter the documents, we selected 12 subdomains that we consider to be

4 https://op.europa.eu/en/web/cellar/home

5 https://op.europa.eu/en/web/eu-vocabularies

Document Type	Moral	Nonmoral	Overlap	Total
Preparatory Act	13,820	7,547	3,352	24,719
International Agreement	1,375	963	409	2,747
Decision	3,991	1,978	775	6,744
Implementing Decision	484	231	199	914
Regulation	2,062	3,188	508	5,758
Implementing Regulation	521	537	185	1,243
Treaty	3	7	0	10
Total	22,256	14,451	5,428	42,135

Table 3: Number of retrieved EU documents for each document type and domain type. Each document can have multiple assigned subjects, which we divided into moral and nonmoral subjects i.e. two domains. The overlap is created, because documents can have subjects from both domains. The ratio between documents in the moral domain and those in the nonmoral domain roughly matches the ratio between the respective numbers of subdomains (12 / 8).

especially moral and 8 subdomains that we consider to be nonmoral. In the following, we list all 20 subdomains:

Moral subdomains:

- criminal law
- rights and freedoms
- family
- migration
- demography and population
- social framework
- social affairs
- culture and religion
- social protection
- health
- organization of work and working conditions
- environmental policy

Nonmoral subdomains:

- political framework
- political party
- monetary economics
- accounting
- maritime and inland waterway transport
- technology and technical regulations
- chemistry
- mechanical engineering

In Cellar, each document is assigned to at least one concept from EuroVoc. Since a document can be assigned to multiple concepts, there can also be an overlap between the subdomains, and therefore some documents will both be considered moral and nonmoral by our categorization.

Cellar also contains documents from the organizations preceding the EU, which was formerly established 1 November 1993. Therefore, we limit the extraction process to documents that are not older than 1994. The remaining amount of extracted documents is 42,135 (extracted on May 12 2025).

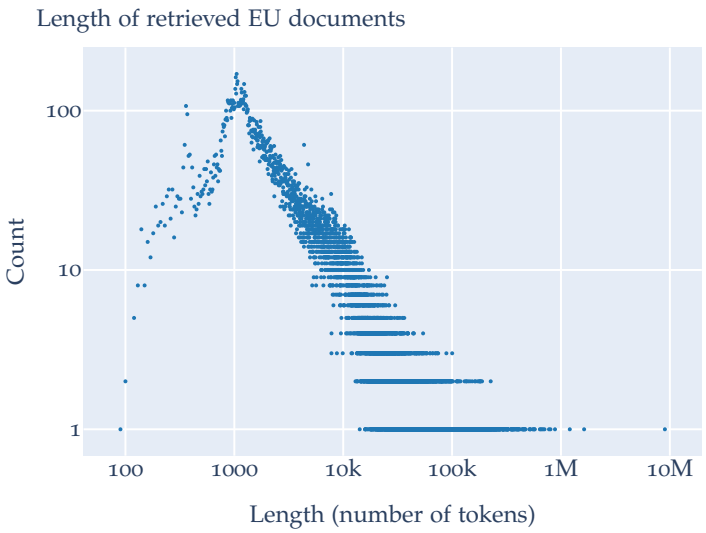

Figure 2: Distribution of EU documents lengths measured in tokens. Each dot represents a range of 5 tokens. It is visible that most documents far exceed the 512 token context window of BERT.

In Table 3 we summarize the distribution of the retrieved documents over document types and the moral/nonmoral domains. It is notable that there are relatively many *preparatory acts* compared to the other documents. The reason for this is that each passed law goes through several iterations, each of which corresponds to a preparatory act. This category also includes staff working documents. We include these documents because we assume that they generally express morality more directly compared to the passed law documents, as these are usually compromises between many different actors.

The low number of treaty documents is not surprising, as there are only four treaties currently in force:[6]

- Treaty on European Union
- Treaty on the Functioning of the European Union
- Treaty establishing the European Atomic Energy Community
- Charter of Fundamental Rights of the European Union

The 10 retrieved treaty documents all amend the Treaty on European Union.

The distribution of document lengths is visualized in Figure 2. Almost all documents (88.7%) exceed the BERT context window of 512 tokens (Devlin et al., 2019), and 208 documents are longer than 100,000 tokens. EU documents can therefore be considered as extremely long documents, making them interesting to study in many contexts including moral foundation prediction.

6 https://eur-lex.europa.eu/collection/eu-law/treaties/treaties-force.html

3.2.2 Annotation corpus

The goal of this work is to develop a model that can extract the moral foundations of EU documents without the access to a dataset of EU documents with moral foundation labels. To the best of our knowledge, such a dataset does not exist currently. To train the model, we therefore use an unsupervised domain transfer technique as presented in Section 4.4.

However, we still need to evaluate the performance of the model on EU documents in addition to the labeled data used for training. For this reason, we create a small corpus of 114 EU documents that are hand annotated for the virtues and vices of the five original moral foundations, i.e. 10 possible labels in total.

When selecting documents for annotation, we have two goals. First, the annotation task has to be practical. If the task is too difficult, the quality of the annotations will suffer and the amount of completed annotations in a limited time will be lower. Second, we need to account for the class imbalance in moral foundations (see Section 3.3). Some moral foundations are more prevalent than others, and positive samples are generally rare.

To make the annotation task practical, we need to limit the lengths of the documents the annotators have to read, since EU documents tend to be extremely long, as we have seen in the previous section. Therefore, we cut off the documents at the next end of a sentence after 1,024 tokens. We use the statistical sentence segmenter[7] from the `spaCy` python package (Honnibal et al., 2020) to determine sentence endings.

Cutting off the documents removes potentially relevant context and signals of morality late in the document. However, we assume that EU documents tend to be long because they cover a lot of details about one subject and not because they cover a lot of different subjects. Our experiences with the annotation work confirm this assumption. Usually, one can get a good idea of what the document is about by reading the title and the first few paragraphs.

Some documents are so long that this procedure cuts them off before the end of the table of contents. Since there would be no content in these cases, we remove all documents that contain a table of contents. These documents would be difficult to grasp for annotators based on the first few paragraphs, as they usually contain many chapters.

Next, we remove documents that are shorter than 512 tokens. Most of these shorter documents are corrigenda, summaries, or amendments. In each case, we are more interested in the original documents. Furthermore, EU documents contain a lot of repetitions and long names/descriptions, which lengthen the document without adding more content. Therefore, these documents only have a few sentences that are actually relevant for the annotation task.

7 `en_core_web_sm` model (version 3.8.0)

At this point, 35,330 documents are left to select a subset for annotation. To further narrow down, we compute the pairwise similarities of the documents using `spaCy` (Honnibal et al., 2020). The document similarity is calculated as the cosine similarity between the averaged word vectors[8] in each document. We remove documents that have a similarity of 0.99 or higher to another document in the corpus. These high similarities occur because some laws consist of multiple documents or go through several stages, each of which is published. By keeping only one of these documents, we ensure that the final annotation corpus does not contain multiple documents that are almost the same.

We are aiming for a relatively balanced dataset and also want to annotate documents that have a high likelihood of expressing morality. We use the document scoring method of the eMFD (Hopp et al., 2021; see Section 2.2) as an estimator for the probability that a document is expressing a certain moral foundation. The eMFD consists of 3,270 words, and for each word the probabilities that it expresses each moral foundation. To compute a document score, the probabilities of all words in a document are averaged. This is by far not a perfect measure since a dictionary does not take context into account, and the eMFD was generated using news texts which are very different from EU documents. However, we do not rely on the eMFD scores being correct. We just need a rough measure for moral foundations as a preselection for annotation, which will then lead to more accurate labels.

The eMFD gives probabilities for the five original moral foundations, not distinguishing virtues and vices. For each foundation, we rank the documents by the respective probability. From the ranking, we select the top k documents for each of the five moral foundations and add them to the annotation corpus. During the annotation process, k was gradually increased according to available resources. The final value was $k = 35$, yielding 114 documents as there is some overlap between the rankings of the different foundations. Furthermore, we selected the top two documents from each foundation as a set of 10 fixed documents that were mandatory for all annotators. The rest of the documents were each annotated by one annotator.

3.2.3 Annotation process

The annotators were provided with a short explanation of moral foundations theory and brief descriptions of each label (virtues and vices of each foundation), including some general examples. Each annotator could choose from all 10 labels. We informed the annotators that documents can express any amount of labels, including no labels at all or contrary ones (i.e. care and harm).

In the instructions, we explained that the documents are cut off at some point. So annotators were aware that they usually did not see the full document. We also asked

8 We use the vectors from the `en_core_web_lg` model (version 3.8.0).

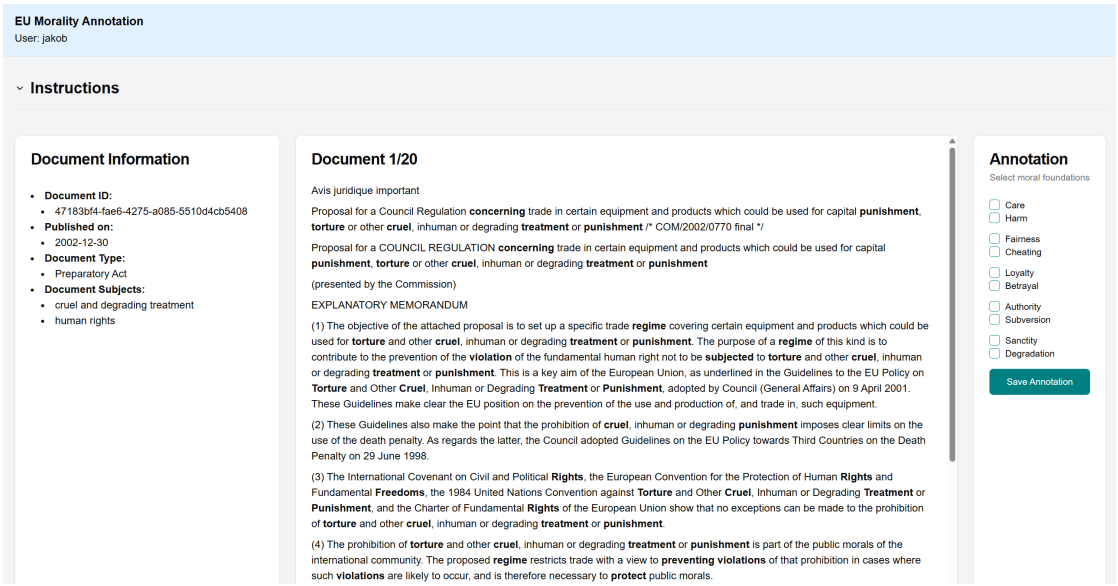

Figure 3: Screenshot of the annotation interface.

them to use as little context knowledge as possible. In Appendix A, we provide the full instructions that were shown to the annotators.

To make the annotation task easier, we provided the publishing date, the document type, and the EuroVoc subjects of each displayed document. Furthermore, we highlighted words that have a score greater than one standard deviation over the mean in any foundation according to the eMFD. Annotators were informed about the highlighting process and its potential weaknesses.

The first 10 documents were the same for each annotator, but their order was randomized to avoid capturing learning effects of the annotators. After finishing the first 10 documents, the annotators were shown random previously unlabeled documents. We asked annotators to complete at least 20 documents (10 fixed, 10 new).

Figure 3 is a screenshot of the annotation interface that was used by all annotators. We developed the interface from scratch and publish its code on GitHub.[9]

3.2.4 Annotation results

There were 6 annotators who annotated 111 documents in total. Three documents from the corpus presented above were skipped by the annotators since they were not readable due to extraction errors. In Table 4 we show annotation statistics of the 10 fixed documents that were annotated by all participants. Previous works like the MFTC (Section 3.1.2; Hoover et al., 2020) and the MFRC (Section 3.1.3; Trager et al.,

9 https://github.com/JakobLindscheid/eu-morality-annotator

	Krippendorff's alpha		Number of documents (50% agreement)		Number of annotations	
	virtues & vices	grouped foundations	virtues & vices	grouped foundations	virtues & vices	grouped foundations
Care	0.312	0.502	5	6	25	30
Harm	0.292		2		10	
Fairness	0.359	0.476	4	5	26	30
Cheating	0.205		1		7	
Loyalty	-0.035	0.099	0	0	3	5
Betrayal	-0.017		0		2	
Authority	0.028	0.001	3	3	16	18
Subversion	0.052		0		4	
Purity	0.103	0.262	0	2	3	8
Degradation	0.126		0		6	
Overall	0.311	0.406	–	–	102	91

Table 4: Summary of annotations on the 10 fixed documents that were annotated by all 6 annotators. We provide additional statistics for the case where the virtues and vices of each foundation are considered as the same label. The statistics include Krippendorff's alpha as a measure for inter-rater agreement, the number of documents where at least half of the annotators agree with a certain label and the total number of annotations for each label.

2022) use Fleiss's kappa (Fleiss, 1971) and prevalence- and bias-adjusted Fleiss's kappa (PABAK; Sim and Wright, 2005) to compute the inter-rater agreement. These measures work under the condition that the annotators for each item are randomly sampled from a larger group. Since this is not the case for us, we report Krippendorff's alpha (Krippendorff, 2018), which additionally works for multi-label classification tasks, such as ours.

When considering all labels together as a multi-label task, we observe an inter-rater agreement of 0.311, which is considered fair (Landis and Koch, 1977). This is in line with the observations of previous works (Beiró et al., 2023; Hoover et al., 2020; Trager et al., 2022), who report similar agreement and conclude that detecting moral foundations in text is relatively difficult, also for human annotators.

In Table 4 we also show the total number of annotations for each foundation and the number of assigned document labels based on a 50% agreement between annotators. Additionally, we group the virtues and vices for each moral foundation into a single label and recalculate the shown statistics. This is done to observe whether annotators disagree over the foundations themselves or rather between the virtue and vice of a foundation. For example, given the sentence

	Moral	Nonmoral	Overlap	Total
Care	30	0	3	33
Harm	25	0	5	30
Fairness	37	2	3	42
Cheating	12	0	0	12
Loyalty	5	0	0	5
Betrayal	3	0	1	4
Authority	19	2	3	24
Subversion	8	0	2	10
Purity	3	0	0	3
Degradation	6	0	1	7
No assigned label	14	16	3	33
Number of documents	81	19	11	111

Table 5: Frequencies of EU documents per foundation and domain based on assigned labels by annotators. For documents that were annotated by more than one annotator we only used labels that were assigned by $\geq$ 50% of annotators. Moral foundations were we observed no or almost no inter-rater agreement are shaded in gray.

"We condemn the killing of the homeless woman."

annotators may disagree whether to choose care or harm, but it is relatively clear that one of the two fits this sentence. In fact, we observe a slightly higher agreement when grouping the moral foundations as visible in Table 4.

In general, there is a fair agreement for the care/harm and fairness/cheating foundations, which also correspond to the most annotations. In contrast to that, there is almost no agreement for the authority/subversion foundation, while there are the third-most annotations for that foundation.

In Table 5 we show the number of documents for each foundation and domain (moral/nonmoral subjects) based on the annotation results from all 111 annotated documents. As before, we use a majority vote ($\geq$ 50% of annotators) to determine the labels of the 10 fixed documents that were annotated by all annotators. All other documents were annotated by one annotator. Again, it is visible that the care/harm and fairness/cheating foundations are the most prevalent labels. The authority/subversion foundation is also detected in a substantial number of documents, while the loyalty/betrayal and purity/degradation foundations are quite rare.

Additionally, we can see that the distinction of moral and nonmoral domains has the expected results. Almost all documents from nonmoral subjects were not labeled with any foundation as visible in the second to last row in Table 5.

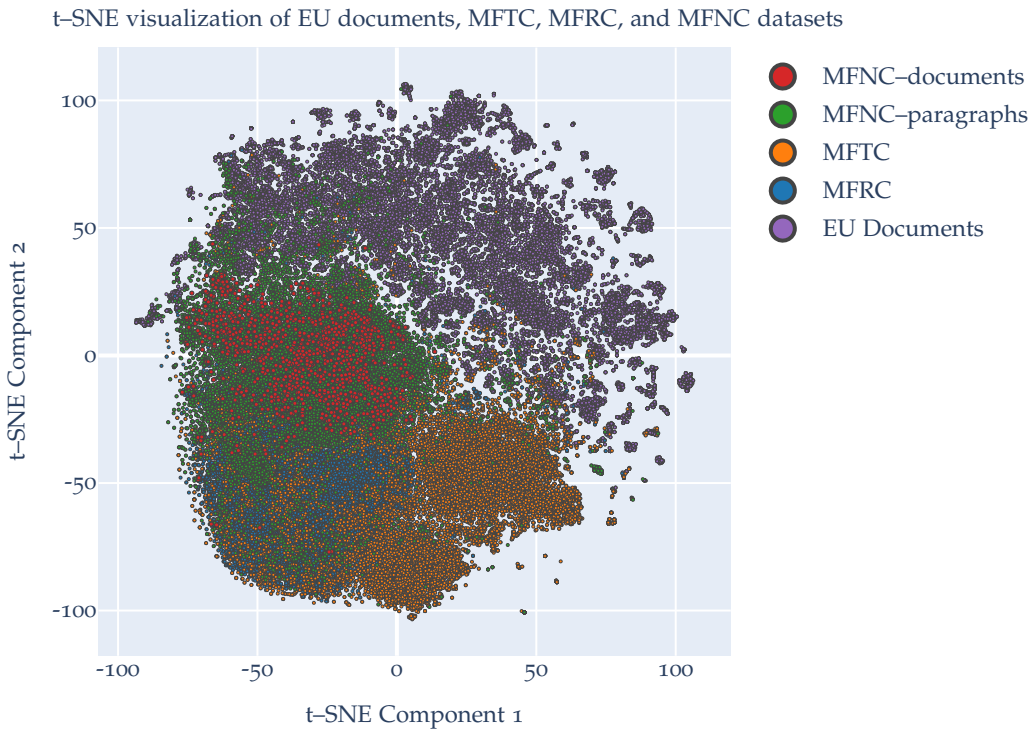

Figure 4: t-SNE visualization of document vectors from the different datasets. The visible separation between datasets, especially EU documents, motivates the use of a domain transfer technique.

3.3 SUMMARY AND EXPLORATORY ANALYSIS

Figure 4 shows 2–dimensional visualizations of the document vectors from the different datasets using t-Distributed Stochastic Neighbor Embeddings (t-SNE; Van der Maaten and Hinton, 2008). The document vectors are generated by averaging the respective token vectors from the `en_core_web_lg` model (version 3.8.0) in the `spaCy` python package (Honnibal et al., 2020).

While there is some overlap between the datasets, it is visible that they are generally located in different parts of the embedding space. Especially, the EU documents are separated quite clearly. As expected, the MFNC–documents are located in the same region as the MFNC–paragraphs. In the MFTC there is some additional separation visible, which may be attributed to the different corpora in that dataset.

Generally, the separation of the different datasets motivates the use of a domain transfer technique. Figure 4 shows that we cannot expect a model trained on one of the datasets to perform well on a different datasets without accounting for that domain transfer.

In Table 6 we show the number of positive samples for each class in addition to the number of documents without any label and the total number of documents for each

	Long Documents				
	Annotated EU documents	MFNC–Documents	MFNC–Paragraphs	MFTC	MFRC
Care	33	644	1,752	3,127	1,821
Harm	30	606	2,695	3,734	2,492
Fairness	42	648	1,884	2,591	2,053
Cheating	12	601	2,367	3,100	2,450
Loyalty	5	634	1,991	2,811	959
Betrayal	4	614	2,191	1,869	764
Authority	24	613	1,856	2,409	1,350
Subversion	10	622	2,274	2,096	1,504
Purity	3	578	1,495	1,307	559
Degradation	7	551	1,926	1,429	1,021
No label	33	0	9,693	7,614	9,892
Dataset size	111	990	20,427	17,990	17,886

Table 6: Number of positive samples for each moral foundation and total number of documents per dataset. Additionally, we show the number of documents without any label.

dataset. In the datasets that do not contain longer documents, we observe a substantial class imbalance between the positive and negative samples for each label. It ranges from 4.3% of documents being positive samples (MFRC, Betrayal) to 20.8% (MFTC, Harm).

It is also visible that the longer texts in the MFNC–documents dataset generally cover more different moral foundations than shorter texts. The annotated EU documents cannot be properly compared to the other datasets because of the low sample size. Some moral foundations are very rare, while others are observed quite often compared to the regular datasets.

4 METHODS

To present our methodology, we first cover some preliminaries, including a task definition in Section 4.1. The presented cross-domain moral foundation classification task comes with several challenges that are approached by different components in our methodology, which will be presented in the following sections. Detecting moral foundations in text is discussed in Section 4.2. EU documents tend to be relatively long (see Section 3.2) compared to most text classification tasks, which is a challenge for most classifiers. In Section 4.3, we present several possible solutions to this challenge. Finally, the domain transfer component of our approach is described in Section 4.4.

4.1 PRELIMINARIES

Before presenting the different methodologies, we introduce the task that needs to be solved by those methodologies. The used notation largely follows Ganin et al. (2016).

We define the cross-domain moral foundation classification task with input space X, representing text documents, and set of labels $Y = \{l_1, l_2 \ldots, l_m\}$, corresponding to the moral foundations with $m = 10$ since we are interested in each of the five original moral foundations including their virtues (i.e. care, fairness) and vices (i.e. harm, cheating). Each document $d \in X$ may be associated with multiple labels from Y or no labels if the document is considered nonmoral. The output for a given instance is therefore a subset of labels $Y_d \subseteq Y$ and the goal is to learn a function $f : X \to 2^Y$.

We consider two different distributions over $X \times 2^Y$: the source domain $\mathcal{D}_S$ and the target domain $\mathcal{D}_T$. The learning algorithm is provided with a labeled source sample S drawn i.i.d. from $\mathcal{D}_S$ and an unlabeled target sample T drawn i.i.d. from $\mathcal{D}_T^X$, which is the marginal distribution of $\mathcal{D}_T$ over X.

$$S = \{(d_i, y_i)\}_{i=1}^{n} \sim (\mathcal{D}_S)^n \; ; \quad T = \{d_i\}_{i=n+1}^{N} \sim (\mathcal{D}_T^X)^{n'}$$

with the total number of samples $N = n + n'$. We are interested in the cross-domain performance of a trained classifier f, which can be evaluated using a labeled sample drawn i.i.d. from $\mathcal{D}_T$.

For this work, the source domain $\mathcal{D}_S$ contains the labeled moral foundations data described in Section 3.1 and the target domain $\mathcal{D}_T$ refers to the EU documents described in Section 3.2. However, since only a very limited amount of labeled EU documents is available to us, we use the MFNC–documents dataset as $\mathcal{D}_T$ to develop the model.

This way, we can properly evaluate and compare different approaches to then decide which model should be used for the EU documents. The MFNC–documents are also considered long documents and can therefore imitate the domain transfer to a dataset of longer documents.

4.2 CLASSIFYING MORAL FOUNDATIONS

The cross-domain moral foundation classification task, we presented above, is a multi-label classification task, which can be modeled with a single classifier. Specifically, we use the pretrained Transformer encoder BERT (Devlin et al., 2019) to obtain the representation of a special classification token `[CLS]`, which is always added as the first token of each input sequence. For classification tasks, BERT feeds this 768–dimensional representation into a "pooler" consisting of a fully connected layer $W_{\text{pool}} \in \mathbb{R}^{768\times768}$ and a tanh activation function.

Finally, we apply a fully connected output layer $W_{\text{out}} \in \mathbb{R}^{768\times m}$, which is called the classification head. The desired output is given as an m–dimensional binary vector, which can be obtained from the model output by applying a sigmoid function and a threshold. To train the model, we minimize the binary focal loss (T.-Y. Lin et al., 2017) between the output of the sigmoid function and the desired output. Focal loss is an extension of the cross-entropy loss, which is commonly used in classification tasks including moral foundation prediction (Kobbe et al., 2020; Zangari et al., 2025). T.-Y. Lin et al. (2017) developed focal loss to address the extreme class imbalance observed in dense object detection by down-weighting the loss for well-classified samples. As we have seen in Section 3.3, moral foundation classification also exhibits substantial class imbalance. The binary focal loss we employ for multi-label moral foundation classification is given by

$$L_{\text{MF}} = \frac{1}{m}\sum_{c}^{m} -w_c(1-p_t)^{\gamma}\log p_t$$

where m is the number of labels, w_c is the weight assigned to the respective class c, γ is a hyperparameter of focal loss and p_t is given by

$$p_t = \begin{cases} \sigma(x_c) & \text{if } y_c = 1 \\ 1-\sigma(x_c) & \text{otherwise} \end{cases}$$

with the sigmoid function σ, the desired output for the respective class y_c and the model output x_c for the respective class c.

An alternative approach is to train multiple models that independently detect a subset of moral foundations. Trager et al. (2022), Nguyen et al. (2024), and Preniqi et al. (2024) compare a multi-label approach as described above with an ensemble of

binary classifiers each detecting a single foundation and find that the latter achieves stronger results. We therefore examine this approach using a set of binary classifiers in addition to the multi-label approach.

For this single-label ensemble approach, we use the same architecture as for the multi-label approach and set $m = 2$, i.e. a positive and a negative class for the respective foundation. The final classification is obtained by selecting the higher output of the two. Furthermore, we replace the binary focal loss with the categorical focal loss function:

$$L_{\text{MF}} = -w_c(1 - \sigma(x_c))^\gamma \log \sigma(x_c)$$

where c is the index of the desired class, i.e. $y_c = 1$.

We train one such model for each moral foundation to obtain the full ensemble, which can make 10–dimensional predictions.

4.3 LONG DOCUMENTS

4.3.1 Aggregation of labels

As we have seen in Section 3.2, the EU documents we want to analyze exceed the 512 token context window of BERT (Devlin et al., 2019) and similar models substantially. However, we still want to process the full documents without truncating them to the required length. To enable these models to process the documents, we split the documents into shorter parts that fit into the context window.

This split is straight-forward for the MFNC datasets (see Section 3.1.1), as the boundaries of paragraphs can easily be identified. While the full MFNC–documents are longer than 512 tokens, most of the paragraphs fit into the context window of BERT.

The EU documents generally do not have clear sections that we can use for this approach. Therefore, we first split the EU documents into sentences using the statistical sentence segmenter[1] from the `spaCy` python package (Honnibal et al., 2020). Some singular sentences are longer than 512 tokens, because documents can contain long enumerations that are detected as single sentences. Therefore, we split these long sentences on line breaks. We then combine consecutive sentences until the limit of 512 tokens is reached to create splits of EU documents. In total, we split 42,135 EU documents into 610,213 chunks of 512 tokens or less.

Models with a limited context size can process the full documents by processing each of the corresponding chunks. We then aggregate the binary output vectors, which represent the set of predicted moral foundations, by computing the union of these sets. This is based on the assumption that if a moral foundation is detected in a part of a document, the whole document contains that moral foundation. Essentially, the labels

1 `en_core_web_sm` model (version 3.8.0)

of the MFNC datasets are determined in the same way as we label each paragraph/-document with the union of annotations contained in them.

A weakness of this approach is that it might disregard context. For example, a document that condemns certain harmful actions will in some parts describe these actions. These parts will be classified as harm, while this label might be incorrect for the whole document.

4.3.2 Longformer

An alternative to the approach presented above is to replace the BERT model with a model that supports a larger context window. Over the last years, several methods have been developed to overcome the Transformer's quadratic scaling of computational and memory requirements with respect to the input size (Tsirmpas et al., 2024). In this thesis, we focus on Longformer (Beltagy et al., 2020), which was one of the earliest approaches but still achieves competitive results in recent studies. Furthermore, Longformer is a relatively robust model that only requires minimal hyperparameter tuning (Dai et al., 2022).

The context window of Longformer has a size of 4,096 tokens, which fits all MFNC–documents and 65.1% of the extracted EU documents. For the remaining documents, the aggregation approach can be applied. In this case, only very few splits per document are necessary. For example, four chunks of 4,096 tokens would be enough to process over 90% of all considered EU documents.

Longformer achieves a linear asymptotic complexity with respect to the input size by employing a dilated sliding attention window (Beltagy et al., 2020). Instead of comparing each token to each other token ($\mathcal{O}(n^2)$), only the w surrounding tokens are taken into account, which reduces the computational cost to $\mathcal{O}(w \times n)$. Stacking multiple of these layers increases the receptive field gradually, similar to convolutional networks (CNN). Assuming w is fixed for all l layers, the final layer has a receptive field of $l \times w$ tokens.

Additionally, this sliding window of size w is dilated with a factor d. This further increases the receptive field without increasing the computational costs by adding gaps of size d to the attention window. The receptive field becomes $l \times w \times d$ when w and d are fixed for all layers. In reality, Longformer uses different settings of d for each attention head, since Beltagy et al. (2020) found that the performance can be increased by focusing some heads on local and others on global context.

In total, this sparse attention approach enables Longformer to process very long sequences efficiently. However, for some tasks, like classification, it is necessary to compute a representation of the whole sequence. To account for this issue, Beltagy et al. (2020) add global attention to some preselected tokens like the `[CLS]` token in classification tasks. With this addition, Longformer can be used as a drop-in replacement for BERT in classification tasks such as moral foundation prediction.

In our work, we treat Longformer as an alternative feature extractor to BERT, while keeping the remaining approach the same.

4.4 DOMAIN TRANSFER

As described in Section 4.1, one core challenge of this work is transferring the knowledge learned from the labeled moral foundations data to the unlabeled EU documents. Fundamentally, a model trained on data from the source domain $\mathcal{D}_S$ can be applied to data from the target domain $\mathcal{D}_T$ since in both cases the input is text and the expected kind of output is also the same. However, this approach may lead to the model learning features that are specific to $\mathcal{D}_S$, which weakens the models performance on $\mathcal{D}_T$. Liscio et al. (2022) show this effect between the different topics covered by the MFTC.

To approach this challenge, Guo et al. (2023) develop a Domain Adapting Moral Foundation inference model (DAMF), which is based on the domain-adversarial training technique introduced by Ganin et al. (2016). Guo et al. (2023) find that DAMF outperforms the BERT baseline in the domain transfer setting and Preniqi et al. (2024) confirm that finding on a slightly different collection of datasets with their MoralBERT model, which is based on DAMF.

For domain-adversarial training (Ganin et al., 2016), a standard model is split into a feature extractor and a label predictor. In addition, a domain classifier is connected to the feature extractor through a *gradient reversal layer*, which multiplies the gradient with a negative constant, but leaves the values unchanged on a forward pass. This means that the feature extractor is trained in a way that minimizes the label prediction loss and maximizes the domain classification loss, which results in domain-invariant features. When no labels are available for data from $\mathcal{D}_T$, the label prediction loss is zero, but the domain classifier is still trained.

In DAMF (Guo et al., 2023) the feature extractor is a BERT model and both the label predictor and domain classifier are classification heads as described in Section 4.2. Furthermore, three new components are added: a weighted loss function, a reconstruction module, and a domain-invariant transformation.

WEIGHTED LOSS FUNCTION The weighted loss function introduces class weights w_c to balance the number of positive and negative samples in the dataset:

$$w_c = \frac{\text{\# negative samples in } c}{\text{\# positive samples in } c}$$

for each moral foundation c.

Guo et al. (2023) develop DAMF for the multi-label approach to moral foundation prediction and we adopt the above class weights for this approach. Preniqi et al. (2024)

introduce the ensemble of single-label classifiers as presented in Section 4.2 and suggest the following class weights for the positive and negative classes:

$$w_{\text{pos}} = \frac{\text{\# total samples}}{\text{\# positive samples in } c}; \quad w_{\text{neg}} = \frac{\text{\# total samples}}{\text{\# negative samples in } c}$$

for the moral foundation c of the respective single-label model. We take the same approach for our single-label ensemble approach.

RECONSTRUCTION MODULE The reconstruction module is added to prevent the BERT encoder from losing too much information in the attempt to maximize the domain classification loss. To keep the generated representations from the BERT encoder close to the original BERT embeddings, a linear layer $W_{\text{rec}} \in \mathbb{R}^{768\times768}$ is added. The reconstruction loss L_{rec} is computed as the mean squared error between the original BERT embeddings and the embeddings reconstructed by the added linear layer.

DOMAIN–INVARIANT TRANSFORMATION Finally, the domain-invariant transformation is an added linear layer $W_{\text{inv}} \in \mathbb{R}^{768\times768}$ after the feature extractor. This layer is supposed to remove any remaining domain-specific information from the embeddings. To still retain text information, the layer is regularized to the identity I with the loss term L_{inv}:

$$L_{\text{inv}} = ||W_{inv} - I||^2.$$

This idea is based on the work by Zhang et al. (2017), where instead of a BERT encoder the embeddings are obtained by encoding sentences with a convolutional model and then weighting them based on their relevance to the respective domain. Therefore, domain-specific information is always retained, which then needs to be removed by the aforementioned transformation. In DAMF the BERT feature extractor is already trained to not retain any domain-specific features. Furthermore, a BERT model ends in a linear layer similar to the one added as domain-invariant transformation in DAMF. It is therefore not entirely clear why this component was added. In Section 5 we ablate all components presented above to evaluate how they are contributing the overall performance.

In total, DAMF consists of four loss terms: the label prediction loss L_{MF} (cross-entropy / focal loss), the reconstruction loss L_{rec} (mean squared error), the domain-invariant transformation loss L_{inv} (identity regularization), and the adversarial domain classification loss L_{adv} (cross-entropy). The weight of these loss terms has to be controlled, since the latter three terms each encourage or discourage the removal of information. Following Guo et al. (2023), we introduce the parameters λ_{adv}, λ_{rec} and λ_{inv} to arrive at the full loss function

$$L = L_{\text{MF}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{inv}} L_{\text{inv}}.$$

Note that we add L_{adv} as a positive term, but through the gradient reversal layer it will effectively be negative for the feature extractor. Ganin et al. (2016) and Guo et al. (2023) control λ_{adv} indirectly through a hyperparameter ν:

$$\lambda_{\text{adv}} = \frac{2}{1 + \exp(-\nu \cdot p)} - 1$$

where p is the ratio between completed and total training epochs. Therefore, λ_{adv} is gradually increased during training. The parameters λ_{rec} and λ_{inv} are directly considered hyperparameters and tuned together with ν.

5 EXPERIMENTS AND RESULTS

5.1 EXPERIMENTAL SETUP

To recap, we are using five datasets in total: 42,135 EU documents, the MFTC (Tweets), the MFRC (Reddit posts), the MFNC–documents (news texts) and the MFNC–paragraphs, which are the same texts as the MFNC–documents, but split into their paragraphs. 111 EU documents are annotated for evaluation, but the EU documents are otherwise unlabeled. Our main experiment is the domain transfer from the labeled datasets (MFTC, MFRC, MFNC) to the unlabeled EU documents (Section 5.9). However, since we only have a very limited number of annotated EU documents available for evaluation, we first tune, validate, and evaluate our approach by treating the MFNC as cross-domain data. Specifically, we use the MFNC–paragraphs without their labels during training, but then use the labels for evaluation. Additionally, we evaluate the performance of the models on long documents by employing the approaches presented in Section 4.3 for the MFNC–documents.

We therefore evaluate each model on three different levels: in-domain (MFTC & MFRC), MFNC–paragraphs (cross-domain) and the longer MFNC–documents (cross-domain). To measure the performance of the models, we use the macro averaged F1 scores over all 10 moral foundations. For the multi-label approach, this is the average of the F1 scores for all classes, while for the single-label ensemble approach, it is the average of the F1 scores for the positive classes of each classifier. The labeled datasets are split into a train (80%), a validation (10%) and a test set (10%).

We clean all datasets in the same procedure as previous works (Guo et al., 2023; Preniqi et al., 2024) by removing hashtags, replacing all mentions with "@user", removing URLs, substituting emojis with textual descriptions and removing non-ASCII characters. In the MFTC and MFRC, previous works furthermore only use labels with at least 50% agreement between annotators. While this is supposed to improve the quality of the labels, it also further increases class imbalance. Averaged over all foundations, only 1.9% of labels would be positive instead of 13.6% in the MFTC and in the MFRC it would be 1.4% instead of 8.4%. Since in our view, class imbalance is a fundamental challenge of moral foundation prediction, we did not apply this rule of agreement. Preliminary tests also confirmed that applying this rule would have a negative impact on performance, and we present these results in Appendix B.

Parameter	single-label ensemble approach	multi-label approach
ν	$[0.1, \mathbf{1.0}, 10.0]$	$[\mathbf{0.1}, 1.0, 10.0]$
λ_{rec}	$[\mathbf{0.1}, 0.5, 1.0]$	$[\mathbf{0.1}, 0.5, 1.0]$
λ_{trans}	$[\mathbf{0.01}, 0.1, 1.0]$	$[0.01, \mathbf{0.1}, 1.0]$
Learning rate	$[5 \times 10^{-6}, \mathbf{1 \times 10^{-5}}, 5 \times 10^{-5}]$	$[5 \times 10^{-6}, \mathbf{1 \times 10^{-5}}, 5 \times 10^{-5}]$
Focal loss parameter γ	$[0.0, 1.0, \mathbf{2.0}, 5.0]$	$[0.0, 1.0, 2.0, \mathbf{5.0}]$

Table 7: Tuned hyperparameters for both approaches. We first conduct a grid search on the loss weight parameters (ν, λ_{rec} and λ_{inv}) and then tune the learning rate and γ separately.

We implement all models and training pipelines in the `huggingface` ecosystem (Wolf et al., 2020). All code and data is published on GitHub[1] and trained models are available on huggingface.[2]

We also employ `huggingface` to retrieve the BERT and Longformer pretrained Transformer models. Specifically, we use the `google-bert/bert-base-uncased` BERT implementation (Devlin et al., 2019) and the `allenai/longformer-base-4096` Longformer implementation (Beltagy et al., 2020).

All of our experiments are performed using compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. In terms of hardware, we mostly use NVIDIA A100 GPUs configured as Multi-Instance GPUs (MIG). Each model is either trained with a 4g.40GB or a 3g.40GB instance. Note that the single-label ensemble approach consists of 10 separate models which can be trained in parallel when using more than one instance. The available memory of 40 GB allows us to use batch sizes up to 64 on the considered datasets. If not specified otherwise, we use a batch size of 64 for all experiments. In Section 5.4, we examine the performance impact of specific batch sizes.

A small number of experiments are run on PNY GeForce RTX 2080TI or NVIDIA L4 GPUs depending on the required resources and availability. Since these GPUs have less available memory, we use gradient accumulation to ensure that the same batch size is used across all experiments. We also use gradient accumulation for some Longformer settings with higher memory requirements. All models that use the MFNC as target domain are trained for 5 epochs. In the experiments with EU documents, we train for 10 epochs to account for the more challenging domain transfer.

	In-domain		Cross-domain			
	MFTC & MFRC		MFNC–paragraphs		MFNC–documents	
	single-label ensemble	multi-label	single-label ensemble	multi-label	single-label ensemble	multi-label
Care	**0.559 ± 0.015**	0.475 ± 0.072	**0.268 ± 0.033**	0.140 ± 0.072	**0.705 ± 0.026**	0.552 ± 0.280
Harm	**0.615 ± 0.009**	0.436 ± 0.077	**0.414 ± 0.036**	0.241 ± 0.028	**0.728 ± 0.037**	0.712 ± 0.079
Fairness	**0.477 ± 0.016**	0.348 ± 0.047	**0.269 ± 0.020**	0.161 ± 0.044	**0.752 ± 0.044**	0.672 ± 0.177
Cheating	**0.555 ± 0.016**	0.372 ± 0.087	**0.367 ± 0.028**	0.206 ± 0.012	**0.760 ± 0.027**	0.733 ± 0.093
Loyalty	**0.524 ± 0.033**	0.352 ± 0.060	**0.257 ± 0.027**	0.164 ± 0.051	**0.739 ± 0.030**	0.671 ± 0.251
Betrayal	**0.367 ± 0.022**	0.062 ± 0.064	**0.220 ± 0.037**	0.150 ± 0.076	**0.682 ± 0.059**	0.607 ± 0.274
Authority	**0.411 ± 0.014**	0.110 ± 0.107	**0.226 ± 0.024**	0.146 ± 0.053	**0.749 ± 0.051**	0.603 ± 0.217
Subversion	**0.481 ± 0.026**	0.210 ± 0.099	**0.318 ± 0.067**	0.162 ± 0.096	**0.770 ± 0.059**	0.601 ± 0.325
Purity	**0.441 ± 0.012**	0.335 ± 0.074	**0.175 ± 0.040**	0.085 ± 0.058	**0.496 ± 0.106**	0.461 ± 0.367
Degradation	**0.380 ± 0.018**	0.228 ± 0.067	**0.211 ± 0.054**	0.123 ± 0.046	**0.620 ± 0.098**	0.600 ± 0.238
Macro Avg.	**0.481 ± 0.007**	0.293 ± 0.069	**0.273 ± 0.007**	0.158 ± 0.028	**0.700 ± 0.016**	0.621 ± 0.169

Table 8: Comparison of single-label ensemble and multi-label approaches using the tuned parameter settings. We report the averaged F1 scores over 5 repetitions. The single-label ensemble approach performs substantially better in all settings and all moral foundations.

5.2 PARAMETER TUNING

As we have seen in Section 4.4, the domain transfer approach has three tunable hyperparameters (ν, λ_{rec}, and λ_{inv}) to control the strength of the different loss terms. Following Guo et al. (2023), we perform a grid search on ν in $[0.1, 1, 10]$, λ_{rec} in $[0.1, 0.5, 1]$ and λ_{inv} in $[0.01, 0.1, 1]$. For this grid search, we choose a learning rate of 1×10^{-5} and the focal loss parameter $\gamma = 2$. These parameters are tuned separately for the best performing setting from the grid search. We test three different learning rates: $[5 \times 10^{-6}, 1 \times 10^{-5}, 5 \times 10^{-5}]$, and then four different values for γ: $[0.0, 1.0, 2.0, 5.0]$. The setting $\gamma = 0$ is equivalent to standard cross-entropy loss.

We compare different runs using the F1 score on the validation split of the MFNC–documents since our primary focus is the performance on long documents. For this, we use the label aggregation strategy described in Section 4.3.1.

The tuning process described above is carried out for both the single-label ensemble and the multi-label approach. In Table 7 we show the tuned parameters as determined by this process.

At this point, we also compare the two tuned approaches to each other. We repeat the training for the optimal parameter settings five times in total and report the averaged results in Table 8. It is clearly visible that the single-label ensemble approach performs

1 https://github.com/JakobLindscheid/EU-Morality

2 https://huggingface.co/collections/JakobLindscheid/eu-morality-6866557b903e0b021d59929a

Adversarial Training	W_{inv}	L_{inv}	L_{rec}	In-domain	Cross-domain	
				MFTC & MFRC	MFNC–paragraphs	MFNC–documents
○	○	○	○	$\mathbf{0.482 \pm 0.007}$	0.242 ± 0.004	0.663 ± 0.019
○	●	○	○	0.481 ± 0.004	0.237 ± 0.011	0.659 ± 0.031
●	○	○	○	0.457 ± 0.004	0.227 ± 0.059	0.625 ± 0.133
●	○	○	●	0.456 ± 0.009	0.257 ± 0.010	0.686 ± 0.040
●	●	○	○	0.465 ± 0.008	0.231 ± 0.016	0.699 ± 0.022
●	●	●	○	0.471 ± 0.004	0.243 ± 0.020	$\mathbf{0.720 \pm 0.014}$
●	●	●	●	0.481 ± 0.007	$\mathbf{0.273 \pm 0.007}$	0.700 ± 0.016

Table 9: Ablation study of different components in the DAMF architecture (Guo et al., 2023). We report the averaged F1 scores over 5 repetitions. While the domain transfer method inhibits the in-domain prediction performance, it improves cross-domain predictions when at least one other component is added. For the MFNC–paragraphs all components contribute positively, while for the longer MFNC–documents the performance can be improved by removing the reconstruction module including L_{rec}.

substantially better in every setting and foundation. Because of that, we do not see a reason to continue using the multi-label approach, and all of the following experiments will use the single-label ensemble approach.

5.3 ABLATION STUDY OF LOSS COMPONENTS

Next, we evaluate the performance impact of the different components in the DAMF architecture (see Section 4.4). Specifically, we consider the use of a domain adversary and the two additional loss terms L_{rec} and L_{inv}. Since the domain-invariant transformation W_{inv} acts as an extra layer, we also evaluate its impact without the corresponding L_{inv}.

In Table 9 we show the averaged F1 scores that different component combinations achieve on the validation splits over five repetitions. It is visible that domain-adversarial training does not improve the performance of in-domain predictions. The performance on cross-domain data is improved when the domain adversary is combined with the other components. We observe that adding the domain-invariant transformation W_{inv} with L_{inv} improves the performance on the MFNC–paragraphs slightly, but more substantially on the longer MFNC–documents when using the aggregation strategy from Section 4.3.1. In general, we can see that adding the extra layer W_{inv} without L_{inv} slightly improves the performance, but adding L_{inv} improves it even further.

With the reconstruction module including L_{rec} the performance on the MFNC–paragraphs improves substantially, while for the MFNC–documents the performance

		Domain transfer approach	
		Incorrect	Correct
Baseline	Incorrect	14.77% (±0.05)	3.49% (±0.02)
	Correct	8.66% (±0.15)	73.08% (±0.16)

		Domain transfer approach	
		Missed	Found
Baseline	Missed	54.52% (±0.19)	14.68% (±0.13)
	Found	2.62% (±0.02)	28.18% (±0.14)

Table 10: We show a two overviews of how the predictions of the full domain transfer approach and the baseline BERT model overlap. All percentages are averaged over all moral foundations and five repetitions. In the left table, we take all labels into account (positives & negatives) and count the cases where the models identify the labels correctly or incorrectly. For the right table, we only consider positive labels and count how many of them are "found" by the respective model. Generally, we see that there is considerable overlap between the predictions of the two models.

can actually be improved by removing the reconstruction module from the full model. Based on this observation, we remove the reconstruction module from the model for the following experiments, since we are primarily interested in classifying long documents.

Based on the reported metrics, our approach with all components outperforms the baseline BERT approach without domain adversarial training in the domain transfer settings. To provide an additional reference point for this finding, we compute the difference between the outputs of these models on the MFNC–paragraphs. In Table 10, we show a accuracy-based (left) and recall-based (right) overview of how the predictions of the two approaches overlap. All percentages are averaged over all moral foundations and five repetitions. The left table takes all labels into account (positives & negatives) and gives percentages of whether the respective models correctly identify the label. Since positive samples are relatively rare, we only consider positive labels in the right table to determine how many of these samples are found by the models. Generally, we see that there is considerable overlap between the predictions of the models, i.e. they produce the same output in most cases. When focusing on the cases where they are different, we observe that the baseline is slightly ahead when considering all labels, but our domain transfer approach is able to identify a larger percentage of positive samples.

5.4 ABLATION STUDY OF CLASS IMBALANCE SOLUTIONS

As already mentioned in several previous sections, we view class imbalance as a fundamental challenge in moral foundation prediction. Our method therefore addresses this challenge in several ways. First, in adopting the DAMF architecture, we follow the idea of Guo et al. (2023) to use class weights in the loss function, which balance the magnitude of the loss per class. Second, we propose using focal loss (T.-Y. Lin et al., 2017), which automatically down-weights well classified samples.

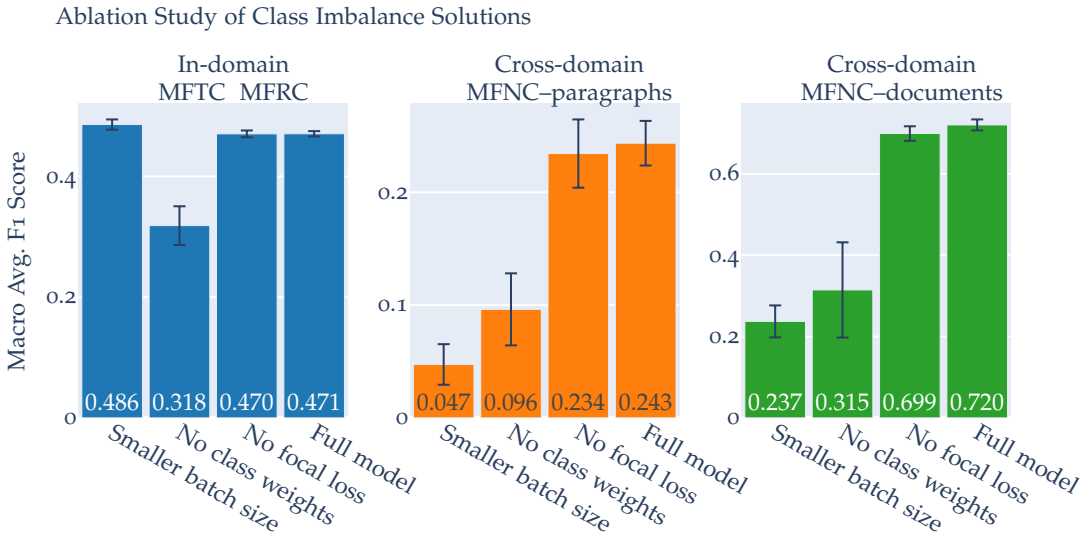

Figure 5: Ablation study of our three approaches to address class imbalance: class weights, focal loss and training with an increased batch size. We report the average F1 score over five repetitions for the full model compared to disabling one component each. The compared batch sizes are 64 (default) and 16 (reduced). In general, we see that each component contributes positively to the performance, though increasing the batch size only improves cross-domain (incl. long documents) performance.

Finally, we consider the invalid sampling problem, which refers to the probability that there are no samples of the minority class in a batch of training data (Hu et al., 2022). For our experiments, where we use the MFTC and MFRC as source domain and the MFNC–paragraphs as target domain, 36.3% of samples are used without labels, as they belong to the target domain and on average 10.6% of the labeled samples are positives. We can estimate the probability of sampling a batch with no positive labels to be 1.1% when using a batch size of 64. Previous works using a similar single-label ensemble approach to ours use a batch size of 16 (Preniqi et al., 2024) or lower (Trager et al., 2022). With a batch size of 16 the probability of invalid sampling becomes 32.7%. Therefore, we suggest training with a higher batch size like 64 to minimize invalid sampling and its performance impact.

In this section, we evaluate the performance impact of each of these three ways of addressing class imbalance. In each setting, we remove one of the approaches while leaving the others unchanged. Removing the class weights is straight forward; the focal loss is disabled by setting $\gamma = 0$ and to examine the impact of an increased batch size, we reduce it to 16 from our default setting of 64.

In Figure 5 we show the averaged F1 scores on the validation splits over five repetitions for each of these settings compared to the full model. It is visible that the impact of using focal loss is relatively minor, but it is still an improvement. Using class weights has a substantial impact in all evaluation settings. Interestingly, the increased

	Aggregation	Longformer	Truncation
Care	$\mathbf{0.718 \pm 0.026}$	0.164 ± 0.047	0.523 ± 0.181
Harm	$\mathbf{0.728 \pm 0.022}$	0.354 ± 0.119	0.590 ± 0.186
Fairness	$\mathbf{0.747 \pm 0.039}$	0.224 ± 0.032	0.589 ± 0.158
Cheating	$\mathbf{0.776 \pm 0.015}$	0.185 ± 0.049	0.458 ± 0.131
Loyalty	$\mathbf{0.754 \pm 0.039}$	0.205 ± 0.115	0.737 ± 0.125
Betrayal	$\mathbf{0.715 \pm 0.057}$	0.061 ± 0.024	0.568 ± 0.301
Authority	$\mathbf{0.743 \pm 0.028}$	0.232 ± 0.104	0.715 ± 0.088
Subversion	$\mathbf{0.767 \pm 0.073}$	0.323 ± 0.039	0.699 ± 0.095
Purity	$\mathbf{0.602 \pm 0.114}$	0.007 ± 0.015	0.421 ± 0.328
Degradation	$\mathbf{0.654 \pm 0.087}$	0.021 ± 0.032	0.462 ± 0.268
Macro Avg.	$\mathbf{0.720 \pm 0.014}$	0.178 ± 0.021	0.576 ± 0.061

Table 11: Comparison of our approaches for moral foundation prediction in long documents. We compare using the union of the predictions over chunks of a document (Aggregation), only applying the model to the first 512 tokens of documents (Truncation) and exchanging the BERT feature extractor with Longformer (Beltagy et al., 2020), which has a context window that fits longer documents. We evaluate the approaches on the MFNC–documents and report the F1 score averaged over five repetitions. Key observations are that the aggregation method performs the best, while the Longformer approach achieves surprisingly bad results even compared to the truncation baseline.

batch size only improves the cross-domain (incl. long documents) performance, while having a slight negative impact on in-domain predictions.

5.5 COMPARISON OF LONG DOCUMENT APPROACHES

As described in Section 4.3, we propose two approaches to handle long documents in moral foundation prediction: aggregating the predictions over chunks of a document and replacing the feature extractor with Longformer (Beltagy et al., 2020), which has a larger context window than BERT (Devlin et al., 2019). To compare these approaches, we use the tuned setting from the previous experiments (i.e. optimal hyperparameters and removed reconstruction module) and only exchange the base model used as feature extractor. Both settings are trained for the domain transfer from the MFTC and MFRC to the MFNC–paragraphs are then evaluated on the MFNC–documents using their different approaches. While the Longformer approach could use the MFNC–documents as target domain data during training as well, we observed in preliminary tests that the performance would suffer substantially in that case. In Appendix C, we provide detailed results of these tests.

Since Longformer has an additional hyperparameter with the attention window size w, we first tune this hyperparameter in the same way we tuned the remaining hyperparameters in Section 5.2. We tune w in $[64, 128, 256]$ and find that $w = 64$ achieves the best performance on the MFNC–documents. This optimal setting is compared to the aggregation approach, as well as the baseline of truncating the documents to their first 512 tokens and then applying the BERT-based model.

In Table 11 we show the averaged F1 scores over five independent training runs for the three methods evaluated on the MFNC–documents. It is visible that the aggregation approach clearly outperforms the other two methods. Surprisingly, the Longformer approach performs substantially worse than aggregation and truncation. We further discuss this observation in Section 6.

5.6 EXCHANGING THE BASE MODEL

In the previous experiment, we used Longformer as a drop-in replacement for the BERT feature extractor in our approach. Of course, we can also use other pretrained Transformer encoders in the same way. To assess the impact of such a replacement, we consider two BERT-related models — RoBERTa (Y. Liu et al., 2019) and ELECTRA (Clark et al., 2020) — that improve the original method by Devlin et al. (2019), but retain the same general architecture (e.g. same input/output format, similar number of parameters, same or similar tokenizer).

With RoBERTa, Y. Liu et al. (2019) improve upon BERT by removing the next sentence prediction objective, adopting dynamic masking, employing a larger byte-level Byte-Pair Encoding (BPE) vocabulary, and training on more data with larger mini-batches for more training steps. These changes result in better performance across several NLP benchmarks. RoBERTa has since been used in many works in the NLP field[3] including previous efforts in moral foundation prediction (Mokhberian et al., 2022; Nguyen et al., 2024).

Clark et al. (2020) propose replacing the masked language modeling pre-training objective used for training BERT and RoBERTa with a new more sample-efficient task called "replaced token detection", which they use to train the ELECTRA model. The approach consists of two models: a generator and a discriminator. The generator is a relatively small model trained for masked language modeling to generate plausible alternatives for some tokens in a sequence. The discriminator, which is the final ELECTRA model, is trained to predict whether each token in the input was replaced or not. This pre-training task is more efficient since it is defined over the whole input sequence and outperforms previous approaches when given a similar amount of compute. When it was published, ELECTRA achieved top scores in several NLP bench-

3 As of June 2025 the RoBERTA paper has been cited over 20,000 times.

			BERT	RoBERTa	ELECTRA
In-domain	MFTC & MFRC	P	0.376 ± 0.009	0.364 ± 0.002	$\mathbf{0.379 \pm 0.013}$
		R	0.637 ± 0.020	$\mathbf{0.700 \pm 0.019}$	0.659 ± 0.024
		F1	0.471 ± 0.004	0.477 ± 0.004	$\mathbf{0.479 \pm 0.011}$
Cross-domain	MFNC–paragrahs	P	0.187 ± 0.036	$\mathbf{0.250 \pm 0.016}$	0.248 ± 0.012
		R	$\mathbf{0.635 \pm 0.081}$	0.374 ± 0.025	0.341 ± 0.049
		F1	0.243 ± 0.020	$\mathbf{0.275 \pm 0.010}$	0.248 ± 0.019
	MFNC–documents	P	0.629 ± 0.029	$\mathbf{0.650 \pm 0.017}$	0.649 ± 0.022
		R	$\mathbf{0.878 \pm 0.052}$	0.767 ± 0.023	0.706 ± 0.076
		F1	$\mathbf{0.720 \pm 0.014}$	0.689 ± 0.015	0.638 ± 0.058

Table 12: Impact of replacing the pretrained model used as feature extractor. We compare our default approach (BERT) with RoBERTa (Y. Liu et al., 2019) and ELECTRA (Clark et al., 2020), which are each used as drop-in replacement without changing the remaining architecture. We report the macro averaged precision, recall, and F1 scores over all moral foundations and five repetitions. Generally, using BERT is still optimal in the long document setting (MFNC–documents), but in the other settings the alternative models achieve a higher performance. In the cross-domain settings, we observe an increased recall when using BERT compared to the other models. This observation is further addressed in our analysis of domain transfer success (Section 5.7)

marks and is still considered competitve with the state-of-the-art today (Bucher and Martini, 2024; Garbas et al., 2024; Tan and H. Liu, 2022).

In Table 12, we show the macro averaged precision, recall and F1 scores over all moral foundations and five repetitions when using BERT, RoBERTa or ELECTRA as feature extractor without changing the remaining architecture. For this experiment, we also use the optimal settings identified in the previous experiments and use the validation splits for evaluation. When evaluating on data from the source domain, all three models perform quite similar with ELECTRA slightly ahead. RoBERTa performs the best on the MFNC–paragraphs (cross-domain) and keeping BERT as feature extractor achieves the highest results on the longer MFNC–documents, which are also cross-domain.

In both cross-domain settings, we observe an increased recall when using BERT compared to the other models. This observation is further addressed in the next section, which is our analysis of domain transfer success (Section 5.7).

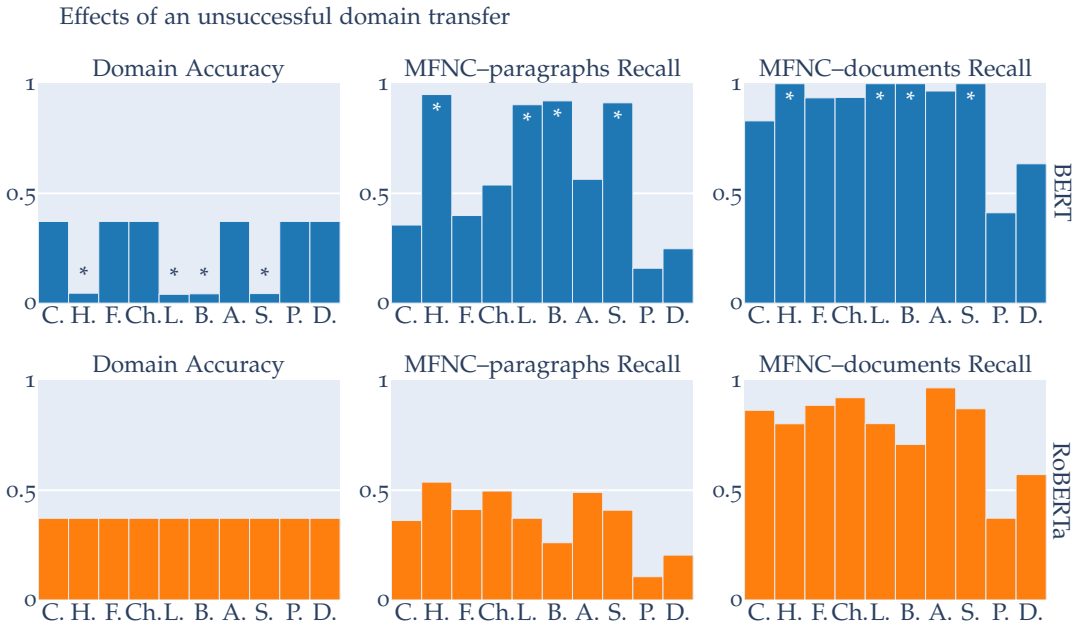

Figure 6: We use the accuracy of the adverary domain classifier to judge the success of the domain transfer. In the left charts, it is visible that the domain transfer fails for some models (marked with *) when using BERT as feature extractor (top charts). With RoBERTa (bottom charts) the domain transfer is always successful. In the center and right chart, we see that a failed domain transfer leads to a substantially increased recall on the MFNC–paragraphs and a recall of 100% on the MFNC–documents (both are cross-domain).

5.7 ANALYSIS OF DOMAIN TRANSFER SUCCESS

To evaluate the performance of our models, we have so far looked at the classification metrics of the label classification head in our architecture. In this section, we will take a closer look at the metrics of the adversarial domain classification head to determine if the domain transfer was successful i.e. the learned features are independent from the domain. In that case, the domain classifier cannot reliably identify the correct domain of a data sample. We expect that the domain classifier becomes either a majority or minority voter, which predicts the same class for all data samples, and its accuracy should reflect the class distribution.

In the left charts in Figure 6, we show the accuracy of the domain classifier. Since the models for each moral foundation are independently trained binary classifiers, the domain transfer may be successful for some, but not for others. Generally, we observe two different accuracy values: $\approx$ 37% indicating a successful domain transfer ($\approx$ 43% of the data is from the target domain) and $\approx$ 4.2% indicating an unsuccessful domain transfer (marked with * in the charts). In the shown results, half of the models using BERT as feature extractor fail at the domain transfer, while all RoBERTa-based models

are successful. We also observe unsuccessful domain transfers when using ELECTRA as feature extractor, but do not show it here to focus on the main points.

In Figure 6, we additionally show the effect of an unsuccessful domain transfer on the performance of the label classification head. Specifically, it leads to a substantially increased recall on the MFNC–paragraphs and a recall of 100% on the MFNC–documents (both are cross-domain). A recall of 100% is not desirable, since a model that always predicts the positive class does not provide any information about the respective documents.

While Figure 6 only shows the results from one training run, we observe the same behavior consistently over five repetitions, which is shown in Appendix D. For BERT, there are between 3 and 6 failed domain transfers per trained ensemble, which always have the effect on recall we described above. We do not observe a connection to specific moral foundations. For "care" and "cheating" we see the lowest amount (1/5) and for "loyalty" the highest amount (4/5) of failed domain transfers. When using RoBERTa as feature extractor, we never observe a failed domain transfer.

Since using BERT as feature extractor achieves a better performance on long documents (see Section 5.6), but with RoBERTa we do not observe failed domain transfers, we will continue using both settings in the following sections.

5.8 COMPARISON TO BASELINES

As a final experiment, before applying our approach to EU documents, we compare its performance to several traditional baselines. Here, we evaluate the approaches on the test split of the MFNC–documents as we are interested in the performance on long documents and the outperformance of Transformers over traditional approaches has been shown both for in-domain and cross-domain classification when texts are not longer than 512 tokens (see Section 2.2). If an approach requires training or tuning, we employ the MFTC and MFRC train splits as our approach is trained in the same way.

The first baseline we consider is the MoralStrength lexicon (Araque et al., 2020), which is an extension of the original Moral Foundations Dictionary (MFD). The MoralStrength lexicon consists of five dictionaries, each representing a virtue/vice pair like care/harm. Each lemma in a dictionary is associated with a value between 1 and 9, where low values represent the vice and high values represent the virtue. To classify a document, the scores of each word in the document are averaged. We determine the final classification by using 5 as a threshold between virtues and vices, i.e. a document with a care/harm score lower than 5 will be classified as harm and a document with a care/harm score higher than 5 will be classified as care. If a document does not contain any word from one of the dictionaries, it is considered a negative for both the respective virtue and vice. This approach requires no tuning and can be directly applied to the MFNC–documents.

	tf-idf-based SVM	eMFD scoring	MoralStrength	Our approach (RoBERTa)	Our approach (BERT)
Care	0.161 ± 0.045	0.018 ± 0.017	0.499 ± 0.035	$\mathbf{0.736 \pm 0.053}$	0.729 ± 0.040
Harm	0.242 ± 0.048	0.297 ± 0.078	0.558 ± 0.073	0.716 ± 0.043	$\mathbf{0.737 \pm 0.042}$
Fairness	0.265 ± 0.047	0.147 ± 0.057	0.700 ± 0.019	0.718 ± 0.050	$\mathbf{0.731 \pm 0.037}$
Cheating	0.212 ± 0.038	0.350 ± 0.080	0.208 ± 0.078	0.755 ± 0.064	$\mathbf{0.769 \pm 0.049}$
Loyalty	0.071 ± 0.045	0.124 ± 0.054	0.736 ± 0.045	0.730 ± 0.045	$\mathbf{0.753 \pm 0.025}$
Betrayal	0.048 ± 0.044	0.311 ± 0.014	0.164 ± 0.075	0.672 ± 0.038	$\mathbf{0.712 \pm 0.075}$
Authority	0.306 ± 0.076	0.228 ± 0.026	0.708 ± 0.024	0.731 ± 0.042	$\mathbf{0.734 \pm 0.034}$
Subversion	0.338 ± 0.048	0.597 ± 0.034	0.267 ± 0.075	0.756 ± 0.048	$\mathbf{0.766 \pm 0.045}$
Purity	0.000 ± 0.000	0.055 ± 0.019	$\mathbf{0.571 \pm 0.028}$	0.497 ± 0.039	0.565 ± 0.114
Degradation	0.007 ± 0.016	0.126 ± 0.062	0.279 ± 0.137	0.529 ± 0.106	$\mathbf{0.662 \pm 0.073}$
Macro Avg.	0.165 ± 0.024	0.225 ± 0.010	0.469 ± 0.018	0.684 ± 0.020	$\mathbf{0.716 \pm 0.019}$

Table 13: Comparison of our approach to three traditional baselines: a SVM with tf-idf features, the eMFD document scoring method and the MoralStrength lexicon (Araque et al., 2020). We report the averaged F1 scores on the MFNC–documents over five repetitions. For all methods, training and tuning was conducted on the MFTC and MFRC. The three baselines are compared to our approach once with BERT and once with RoBERTa as a feature extractor. Both settings generally outperform all three baselines.

For the second baseline, we employ the document scoring method of the eMFD (Hopp et al., 2021; see Section 2.2 & 3.1.1). The eMFD contains probabilities for expressing each virtue/vice pair for over 3,000 words. Next to these five probabilities, each word is assigned five additional values, which are "sentiment scores" for each virtue/vice pair, creating a 10 dimensional vector per word. To compute a document score the vectors of all words in a document are averaged. The probabilities can be used to determine whether a label should be assigned in a certain virtue/vice pair, and the sentiment scores can be used to decide between the virtue and the vice. For this, we need to determine a threshold for each of the 10 values. To find these thresholds, we first compute the document scores on the training data and use the means of these scores as thresholds. Then, we compute the document scores of the MFNC–documents and classify them the following way: If a probability for a virtue/vice pair is lower than the respective threshold, no label is assigned in that pair. Otherwise, we classify it as the virtue if the respective sentiment score is higher than its threshold or as the vice if it is lower than that threshold.

While the previous two baselines are dictionary-based, our final baseline is a Support Vector Machine (SVM; Cortes and Vapnik, 1995) using tf-idf weights as features (Joachims, 1998). To obtain the multi-label output, we train one SVM for each moral foundation, similar to our single-label ensemble approach. The documents are tok-

enized by whitespace and basic punctuation, and from the resulting term frequencies we vectorize the documents using tf-idf. These tf-idf vectors are the inputs for the SVMs. The SVMs use linear kernels and are trained on the train split of the MFTC and MFRC as we described above.

For all three baselines, we conduct five independent repetitions. The impact of randomness is mostly limited to the data splits, but for the SVM-based method it also affects the initial state. In Table 13 we show the averaged F1 scores on the test split for all three baselines and our approach once with BERT and once with RoBERTa as a feature extractor. Both settings of our approach clearly outperform the baselines, with the BERT-based setting performing the best on average. Interestingly, the MoralStrength lexicon performs slightly better than our approach on the purity foundation, which generally seems to be a weak point for our models compared to the other foundations.

5.9 APPLICATION ON EU DATA

With each component of our approach tuned, validated, and evaluated, we can now apply it to EU law documents. In contrast to the previous experiments, we add the MFNC–paragraphs to the source domain together with the MFTC and MFRC, since the EU documents now serve as target domain. We use the label aggregation approach described in Section 4.3.1 to process long documents, i.e. the 42,135 EU documents are split into 610,213 chunks of 512 tokens or less. Using all of these texts for domain-adversarial training would cause the domains to be highly imbalanced, as the three source domain datasets together only contain 56,303 samples. Therefore, we balance the domains by adding a randomly sampled EU document to the training subset until the number of EU document chunks in the subset is greater than or equal to the size of the source domain. With a train split of 80%, the training data finally consists of 45,025 samples from the source domain and 45,034 chunks of EU documents as target domain.

Otherwise, we use the same setting as in the previous experiments (i.e. single-label ensemble with tuned parameters and removed reconstruction module), but train for 10 epochs instead of 5. We also conduct this experiment once with each of BERT and RoBERTa as feature extractors for the reasons shown in Section 5.7.

After training has concluded, we apply the models to the entire set of EU document chunks. This results in two outputs (positive/negative) per moral foundation and model. To make the outputs somewhat comparable, we apply the softmax function and retain the output of the positive class, which is now scaled between 0 and 1, similar to a probability. We then aggregate these outputs for each document by taking the maximum observed value for each moral foundation over all the chunks of the document. This is equivalent to the method described in Section 4.3.1, which first determines the assigned classes to the chunks and then takes the union of the resulting

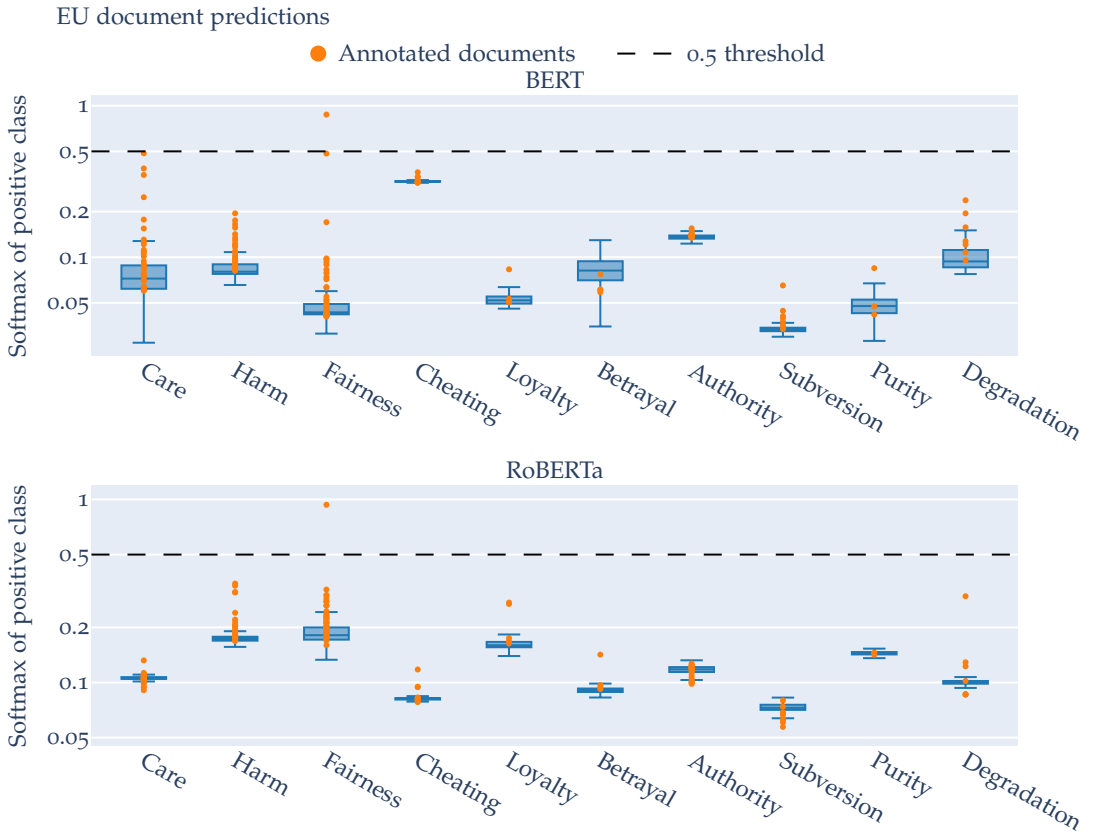

Figure 7: Distributions of model outputs on EU documents after applying the softmax function. We train two models for moral foundation prediction on EU documents, each using a different model as feature extractor (BERT & RoBERTa). To make the charts more interpretable, the y-axis is scaled logarithmically. Instead of boxplot outliers, the orange dots represent the manually annotated documents. Almost all documents are far below the traditional 0.5 threshold, which leads us to consider alternatives in the following analysis.

binary set. Here, we are interested in a closer analysis of the outputs, so we aggregate first to later determine the assigned classes.

In Figure 7, we show boxplots of these aggregated outputs, but instead of adding outliers as individual points, the orange dots represent the annotated EU documents. Usually, we would determine the assigned classes by taking the higher value of the positive and negative outputs, i.e. the positive output has to be higher than 0.5 after applying the softmax function. From the boxplots, we can see that almost all predictions are far below this threshold, and only one annotated document would be classified correctly if this threshold was used.

To still get an idea of how well the models perform on the annotated documents, we use two different methods. First, we compute the Area Under the Receiver Operating Characteristic Curve (ROC AUC), which does not require a fixed threshold. Second,

	Inter-rater agreement	Upper whisker threshold: Precision BERT	Precision RoBERTa	Recall BERT	Recall RoBERTa	F1 score BERT	F1 score RoBERTa	ROC AUC BERT	ROC AUC RoBERTa
Care	0.312	**0.538**	0.263	**0.212**	0.152	**0.304**	0.192	**0.624**	0.384
Harm	0.292	0.522	**0.545**	**0.400**	0.400	0.453	**0.462**	**0.766**	0.696
Fairness	0.359	0.433	**0.500**	**0.310**	0.238	**0.361**	0.323	0.513	**0.597**
Cheating	0.205	**0.300**	0.214	**0.250**	0.250	**0.273**	0.231	**0.524**	0.524
Loyalty	-0.035	0.077	**0.105**	0.200	**0.400**	0.111	**0.167**	0.442	**0.749**
Betrayal	-0.017	0.000	**0.091**	0.000	**0.250**	0.000	**0.133**	0.271	**0.776**
Authority	0.028	**0.333**	0.000	**0.083**	0.000	**0.133**	0.000	**0.618**	0.568
Subversion	0.052	**0.172**	0.000	**0.500**	0.000	**0.256**	0.000	**0.737**	0.171
Purity	0.103	**0.500**	0.000	**0.333**	0.000	**0.400**	0.000	**0.654**	0.386
Degradation	0.126	0.200	**0.333**	**0.429**	0.429	0.273	**0.375**	**0.772**	0.515
Overall	0.311	**0.308**	0.205	**0.272**	0.212	**0.256**	0.188	**0.592**	0.537

Table 14: To evaluate the models trained for EU documents, we employ two different methods since the traditional 0.5 threshold does not work well. First, we use the upper boxplot whisker as an alternative threshold and report standard classification metrics. Second, we compute the Area Under the Receiver Operating Characteristic Curve (ROC AUC) as this does not require a fixed threshold. Additionally, we report Krippendorff's alpha as a measure for inter-rater agreement from Section 5. Moral foundations were we observed no or almost no agreement are shaded in gray. Generally, we see that for each moral foundation at least one of the two models achieves a reasonable performance, especially considering the difficulty of the task.

instead of using the same 0.5 threshold for all labels, we use the upper whiskers (75% percentile + 1.5× interquartile range) of the respective boxplots as thresholds, i.e. the documents traditionally considered outliers in boxplots. This ensures that the threshold are adapted to each output distribution. We decided against using a fixed percentile of outputs as this would lead to the same amount of classified documents for each moral foundation. Analyzing which moral foundations are the most and least represented by the EU would be impossible in that case. Since the upper whisker is computed by adding 1.5 times the interquartile range to the 75% percentile, the number of "outliers" (documents with a softmax score higher than the upper whisker) is different for each boxplot. Like this, we select the documents that were assigned the highest probability of expressing a certain moral foundation, but the number of selected documents depends on the specific distribution of all outputs.

Generally, using the upper boxplot whiskers as thresholds has two main weaknesses. First, this method assumes that positives are "outliers" in the context of a boxplot, which is not necessarily the case for all datasets. Second, the classification of a document depends on the outputs of other documents, i.e. documents can only be classified in the context of a corpus of documents. In our case, the impact of these two points is

minimal. We do not expect many EU documents to contain especially moral content. Therefore, assuming that moral documents are outliers is reasonable. Furthermore, we are not aiming to classify documents without context, as we use the full corpus of extracted EU documents.

In Table 14, we report the ROC AUC and standard classification metrics when using the upper whisker as threshold. We also add Krippendorff's alpha as a measure for inter-rater agreement, which we calculated in Section 3.2.4. The moral foundations for which we observed no or almost no agreement are shaded in gray, and the corresponding classification metrics should only be analyzed in that context.

According to the ROC AUC, there is at least one model per moral foundation that achieves a reasonable performance, while fairness and cheating are the worst performing foundations when considering the better performing model for each label. From the F1 scores, we can again see that at least one model per moral foundation achieves reasonable performance. Here, the models generally perform the worst on the labels that also show almost no inter-rater agreement.

Next, we look at the number of classified documents per moral foundation for both the standard 0.5 threshold and our idea of using the upper whisker as threshold. These results are visible for both models in Figure 8. Generally, we see that fairness is the most classified moral foundation, usually followed by care and harm. Betrayal and authority are observed the least when considering all four settings. We also observe that in most cases more documents are assigned to virtues than to their respective vices. Exceptions to this are the authority/subversion and the purity/degradation pair. When grouping the labels into individualizing (care/harm, fairness/cheating) and binding (loyalty/betrayal, authority/subversion, purity/degradation) moral foundations, substantially more documents are assigned to individualizing than binding foundations in all settings. Between 56.5% and 66.2% of all assigned labels belong to individualizing foundations across all four settings.

By counting the documents that have no assigned label in a setting, we can determine how many documents are classified as "nonmoral". Depending on the setting, between 75.4% (BERT, upper whisker threshold) and 98.9% (RoBERTa, 0.5 threshold) of documents are considered nonmoral. In Section 3.2, we selected especially moral and nonmoral Eurovoc subdomains to decide which EU documents to extract. When accounting for the different amounts of documents from moral and nonmoral domains, we observe that documents from nonmoral domains are more likely to not have any assigned moral foundation. However, this difference is only between 0.34 and 8.1 percentage points e.g. for BERT with the 0.5 threshold 98.33% of nonmoral documents and 97.99% of moral documents are labeled as nonmoral.

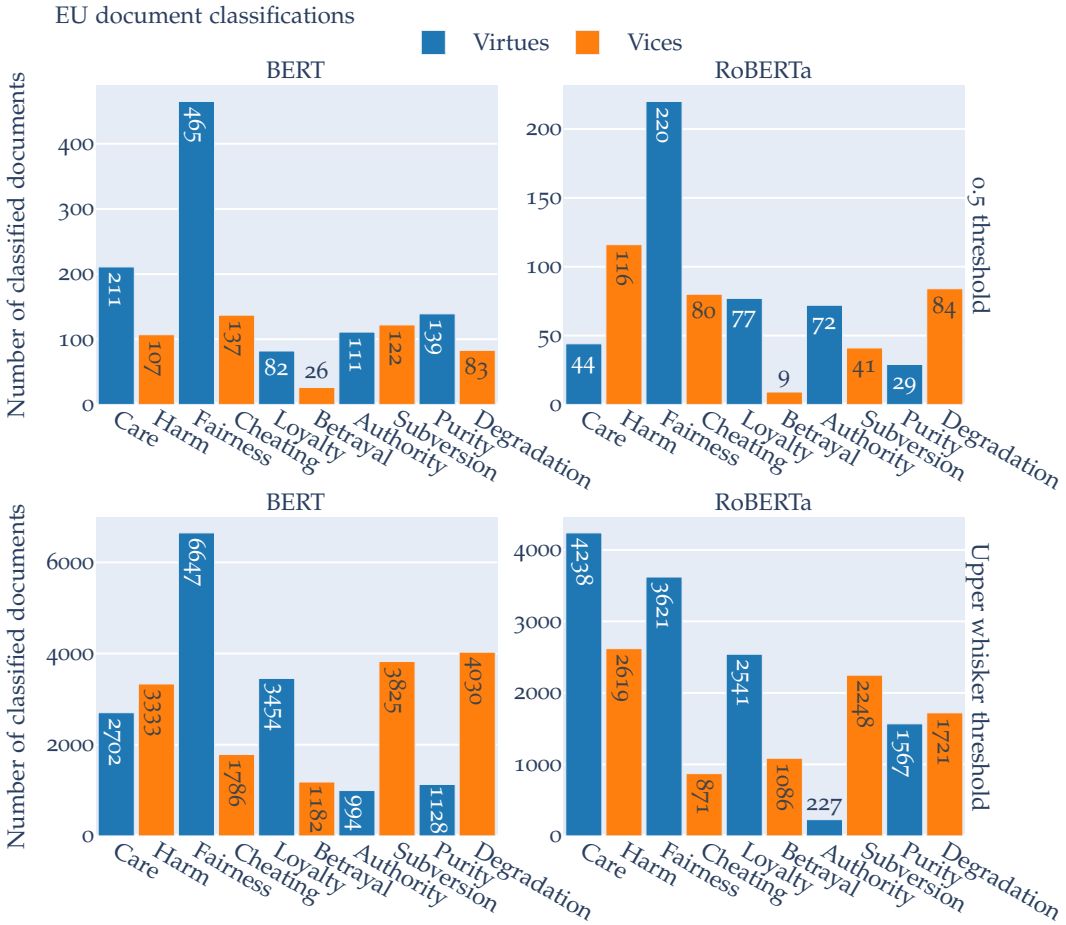

Figure 8: We count the number of classified EU documents for each moral foundation, by applying two different thresholds (0.5 and the upper whisker of the respective boxplot) to the output of the two trained models. The main observations are that fairness is the most prevalent foundation, while betrayal and authority are observed the least. Generally, we see that substantially more documents are assigned to the individualizing foundations (care/harm, fairness/cheating) than to the binding foundations (loyalty/betrayal, authority/subversion, purity/degradation).

5.10 QUALITATIVE ANALYSIS

As we have seen from the lacking inter-rater agreement for some of the labels, determining which moral foundations are expressed by a text can be subjective. The classification metrics reported in the previous section may therefore be insufficient to judge the performance of our models. In this section, we provide an additional reference point by presenting some example documents and the corresponding outputs of our models. To choose example documents, we refer to our analysis of moral diversity in the previous section. Specifically, we select the documents that receive the highest

score in the most prevalent moral foundation (fairness) and the least prevalent moral foundations (authority and betrayal). While it might appear interesting to also present the lowest-scoring documents, these documents usually contain almost no text. For example, the document that receives the lowest fairness score just contains two sentences announcing that a certain agreement is entering into force.

In the following, we will summarize each of the documents and list the assigned labels to these documents. For each document, we add a footnote with a link to the full texts, which are available online. To contain the length of this Section, we will consider the outputs of the model that uses RoBERTa as the feature extractor exemplarily for both settings. We assume that due to the similarities between the two models, our findings translate to the BERT-based model.

The document that receives the highest fairness score is a preparatory act from May 2020.[4] It evaluates the EU's legal framework on equal pay between women and men. In the document, challenges are identified that prevent the corresponding laws from fully working. The Eurovoc concepts that are assigned to this document are the following:

- equal treatment
- gender equality
- position of women
- sexual discrimination
- women's rights
- working conditions

In addition to fairness, the scores for the care and cheating foundations are greater than the respective upper whisker (see Sect. 5.9). Only the scores for fairness and care are greater than 0.5. Clearly, fairness is a reasonable label for this document. Since current shortcomings in this area are also discussed, the moral foundation of cheating also fits. The care foundation often relates to vulnerable or disadvantaged individuals. Most likely, it was assigned to this document because it also discusses victim support.

Next, we present the document that was assigned the highest score for authority.[5] It is also a preparatory act, specifically the chapter on Estonia in the 2024 Rule of Law Report. The report covers the independence and efficiency of the judiciary, anti-corruption measures, media freedom, and institutional checks and balances. This relates to the following Eurovoc concepts:

- corruption
- freedom of the press
- democracy
- rule of law

The moral foundations of authority, fairness, and purity exceed the threshold of 0.5. Since all upper whiskers are lower than that, harm, cheating, and subversion are also

4 http://publications.europa.eu/resource/cellar/02beb5e4-6cb0-11ed-9887-01aa75ed71a1
5 http://publications.europa.eu/resource/cellar/8e9aef8b-4ab1-11ef-acbc-01aa75ed71a1

assigned when using that threshold. From the Eurovoc concepts, it is relatively clear why the former three foundations were assigned. "Rule of law" corresponds to the authority foundation with the document also discussing reforms to strengthen the judicial independence. The report puts an emphasis on equality (e.g. rules applying equally, equal access to legal protection), which fits the "democracy" Eurovoc concept and the fairness foundation. Finally, purity may not be as clear as the other labels, but topics like anti-corruption efforts and institutional integrity can also be seen as part of this moral foundation.

Finally, the document with the highest output for betrayal is an "Implementing Decision" document from June 2014.[6] In the document, three people are added to a list concerning restrictive measures against the Central African Republic. According to the listed reasons, these individuals are responsible for undermining peace and security in the Central African Republic. The document belongs to the Eurovoc subdomain "criminal law". Our model assigns a score greater than 0.5 to the cheating and betrayal foundations for this document. Interestingly, all labels, except care, are assigned a higher score than the respective upper whisker. The reasons for assigning these labels most likely stem from the descriptions of the committed crimes. All three individuals "undermine the peace, stability or security" of the Central African Republic. Betrayal may be assigned because of disloyalty to the own country, but subversion also fits this description. Generally, the document describes many serious crimes including torture, executions, and destabilizing a legitimate government. This is most likely the cause for the variety of assigned moral foundations for this document.

From this analysis, we conclude that our models produce mostly reasonable predictions for EU documents and are able to pick up signals for moral foundations in the texts. Together with the evaluation on the annotated documents in the previous Section, this demonstrates the usefulness of our approach.

6 http://publications.europa.eu/resource/cellar/d14808ea-fb6e-11e3-831f-01aa75ed71a1

6 DISCUSSION

6.1 FINDINGS

From the results of our experiments, there are some relevant findings for the broader field of moral foundation prediction in text documents. In Section 5.2, we find that the single-label ensemble approach substantially outperforms the multi-label approach, which further solidifies the findings by Trager et al. (2022), Nguyen et al. (2024), and Preniqi et al. (2024). This suggests that there are at most weak inter-label dependencies between moral foundations that are not beneficial to model jointly. Training specialized models for each label is more effective, which may also be due to differences in model capacity. The model trained for multi-label classification only has slightly more parameters than a model trained for binary classification as they only differ in the last layer ($W_{\text{out}} \in \mathbb{R}^{768\times10}$ compared to $W_{\text{out}} \in \mathbb{R}^{768\times2}$). Therefore, the single-label ensemble approach has roughly 10 times more available parameters per moral foundation. In our case, inter-label dependencies are clearly not sufficient to compensate for this difference.

From our three tested approaches to class imbalance in Section 5.4, we saw all of them improve the cross-domain classification performance. The class weights suggested by Guo et al. (2023) had a large impact in all settings as expected. Our idea of using focal loss (T.-Y. Lin et al., 2017) also improved the F1 scores slightly. Interestingly, our suggestion of training with an increased batch size only improved cross-domain performance while having a slight negative effect on in-domain performance. Clearly, our initial reasoning for increasing the batch size, i.e. to avoid batches only samples of one label, does not fully explain these observations since we would expect this to improve in-domain classifications as well. Invalid sampling may not be a problem in moral foundation prediction, but increasing the batch size also increases the number of cross-domain samples in a batch, which likely leads to more reliable gradients from the domain classifier and improved domain-invariant feature learning. This would lead to better generalization beyond the source domain, improving cross-domain performance without having a substantial impact on in-domain performance.

6.2 IMPLICATIONS

In our analysis of EU documents, we come to the same conclusion as Grosfeld et al. (2024), confirming their hypothesis that individualizing moral foundations are over-represented in comparison to binding foundations. Next to its economic goals, the primary aim of the EU as a post world war organization is to promote and protect peace. Its values are human dignity, freedom, democracy, equality, rule of law, and human rights.[1] Therefore, it is not surprising that the moral foundations of care and fairness are represented very well in EU law documents. However, this lack of moral diversity may lead to the EU being perceived as less legitimate in the eyes of its citizens. According to Graham, Haidt, and Nosek (2009), individualizing foundations are more related to liberal ideologies, while conservative people are rather guided by binding moral foundations. This is in line with Euroscepticism being more prevalent in right-wing parties (Werts et al., 2013).

Finally, we want to discuss our findings in classifying long documents for moral foundations and their implications for MFT. We compared the label aggregation approach, which combines predictions from smaller parts of a document, with Longformer (Beltagy et al., 2020), which is able to consider a whole document at once. We found that the label aggregation approach substantially outperforms Longformer. We believe that these two approaches represent two fundamentally different ways to label long documents. One can either consider the context of the full document and annotate according to the moral foundations expressed by the document as a whole, or alternatively one can annotate each statement in the text separately and label the document with the moral foundations that are expressed at any point in the document. As an example, consider a document that condemns the actions of an oppressive government. As a whole, such a document may express the moral foundation of care if it generally supports victims, but some statements that describe the events in the country could be related to the moral foundation of harm. In this case, the two labeling approaches lead to different results. In fact, multiple annotators independently asked us which of these two methods to use during our annotation effort. Currently, there is no answer to this question and we urge the research community to further explore how morality is perceived in (long) texts, beyond self-report studies such as most annotation processes.

6.3 LIMITATIONS

In our work, we most likely observe the outperformance of the label aggregation approach, because the dataset we use for evaluation (MFNC–documents) are generated by aggregating all annotations of text portions within a document. This leads to an inherent advantage for an approach that also aggregates labels across shorter text spans.

1 See article 2 of the Lisbon treaty and the EU Charter of Fundamental rights for the values of the EU.

Unfortunately, we were not able to further investigate this, since there are no other datasets available that contain long documents and moral foundation labels.

Another limitation of our work, that is caused by the MFNC–documents dataset, is the changed class balance in that dataset compared to other datasets. While in the MFTC, MFRC and MFNC–paragraphs the negatives are by far the majority class, they become the minority in the MFNC–documents as they are aggregated from the MFNC–paragraphs. Since we do not train any model directly with the MFNC–documents, this does not influence the effectiveness of our approach, but the performance evaluation is affected by the high percentage of positive samples. Specifically, always predicting the positive label becomes a strong baseline when considering the F1 score of the positive class, as we do in this work. With a recall of 100% and a precision of $> 50\%$ the F1 score is at least 66.6%. This is exactly what we observe when the domain transfer is unsuccessful as we show in Section 5.7. Therefore, the F1 score may be inflated for models where the domain transfer is unsuccessful more often. We also assume that this is the reason why the reconstruction module of the DAMF architecture (Guo et al., 2023) improves the performance on the MFNC–paragraphs, but not on the MFNC–documents.

Further limitations include the focus on 10 dimensional moral foundation prediction i.e. grouping the virtues and vices into five labels was not considered, and the limited hyperparameter tuning. While we tuned all hyperparameters, we assumed their independence at several points and also carried the tuned parameters over to changing architectures (exchanged base models, loss component ablations). This was done to keep the computational costs of our experiments realistic, but to improve the validity of our results, more hyperparameter tuning could be done.

As a final limitation, we have to account for the limited size and inter-rater agreement of our dataset of annotated EU documents (see Section 3.2.4). For six of the ten labels we observe a lower inter-rater agreement than expected from previous annotation efforts like the MFTC (Hoover et al., 2020) and MFRC (Trager et al., 2022). Furthermore, four of the labels are assigned to fewer than 10 of the 111 documents. The evaluation results in Section 5.9 should be interpreted in that context.

7 CONCLUSION

In this thesis, our aim was to develop a context-aware approach to determine the moral foundations expressed by EU law documents. This goal comes with several challenges. Since no dataset exists that contains EU documents labeled for moral foundations, we employ domain adversarial training to finetune a Transformer encoder. With this technique, our models are trained on labeled news texts and social media posts, but learn domain-invariant features for moral foundation prediction that transfer relatively well to EU documents. To evaluate our approach, we manually annotate 111 EU law documents with the expressed moral foundations. The performance of our models varies per moral foundation, but in general we find that they perform well considering the difficulty of the task. This difficulty can also be seen in the partly lacking inter-rater agreement in our annotation effort, which demonstrates the subjectiveness of identifying moral foundations. To provide another reference for the performance of our approach, we demonstrate its results on some example documents. From both the metrics on the annotated documents and the qualitative analysis, we conclude that the models were able to pick up a signal in the text to identify moral foundations.

The primary challenge in classifying EU documents is that they tend to be extremely long. We develop two fundamentally different approaches to process long documents ($>$ 512 tokens) that also correspond to different labeling techniques. In the first approach, we split the documents into chunks that fit into the 512 token context window of BERT and BERT-related Transformer encoders. The predicted classes are then aggregated into predictions for a whole document by taking the union of all outputs from that document. An annotator might take a similar approach when assigning labels for moral foundations that are observed only in parts of a document but not necessarily expressed by the document as a whole when putting that part into context. Our second approach replaces the model used as a feature extractor with Longformer, which can process much longer sequences i.e. full documents at once. This would be similar to an annotator taking the full context and intention of a document into account to assign a label. When evaluating our approaches on long news documents that are labeled with the expressed moral foundations, we find that the label aggregation approach substantially outperforms the Longformer approach. However, we also believe both labeling approaches to be reasonable. Exploring which of the two methods is closer to the moral reactions of people reading a text is an interesting avenue for future work in Moral Foundations Theory.

We also find that the label aggregation approach outperforms traditional dictionary-based and term-frequency-based methods when predicting moral foundations for long

documents. Even though the benefits of using a Machine Learning approach for general moral foundation prediction have been shown several times, recent research about MFT still largely employs dictionaries for text analysis. With our work additionally addressing the limitation of the restricted context length in Transformer models, researchers should consider adopting context-aware Machine Learning approaches over traditional dictionary-based methods in future analyses.

While the document length is a challenge specific to the data of interest, we identify the class imbalance between positive and negative samples as a general challenge in moral foundation prediction. We add class weights to the loss function and exchange the standard cross-entropy loss with focal loss to address the class imbalance. This results in a substantially improved performance for the in-domain, standard cross-domain, and long document cross-domain setting. This effect is stronger in the cross-domain settings than for in-domain predictions. Additionally, we suggest training with an increased batch size to avoid the invalid sampling problem. This also improves the performance in both cross-domain settings, but it has a slight negative effect in the in-domain setting. We conclude that class imbalance is a fundamental challenge in moral foundation prediction, which is magnified in domain transfer tasks.

Finally, by applying our approach to EU documents, we confirm previous hypotheses of a lack of moral diversity in the EU's laws and communications. Generally, individualizing moral foundations (care/harm, fairness/cheating) are overrepresented in comparison to binding moral foundations (loyalty/betrayal, authority/subversion, purity/degradation), which usually rather resonate with conservative ideologies. A weak belief in the moral rightness of an organization like the EU may negatively affect its perceived legitimacy. At the same time, liberal democracies depend on being perceived as legitimate by their citizens in order to be effective. The lack of moral diversity we found may therefore be a critical issue for the EU to resolve.

Based on our research, there are many possibilities for future work. As we already discussed above, an interesting question would be in which way reading a text causes moral reactions, specifically for long documents. While this question rather belongs in the field of psychology, computer science can also contribute further by exploring approaches for long documents beyond the ones presented in this thesis. Hierarchical Transformers (Pappagari et al., 2019) for example might be an interesting hybrid between our methods.

Moral Foundation prediction in general and especially in cross-domain settings might benefit from exploring more approaches to the class imbalance challenge. These may include resampling techniques, data augmentation or further loss function adaptations. As with focal loss, many methods have been developed to deal with more extreme class imbalance in other machine learning tasks, which could be used as inspiration.

Generally, we have shown the difficulty of the domain transfer task and the limitations of our approach. Future work could further advance the DAMF architecture to approach the challenges we observed or suggest entirely different domain transfer techniques such as contrastive learning (Fang et al., 2020), continued pretraining on the target domain before finetuning (Gururangan et al., 2020), or employing domain specific base models (e.g. LEGAL-BERT (Chalkidis et al., 2020)).

Finally, our work does not take Large Language Models (LLMs) into consideration. With their impressive zero-shot classification performance and usually longer context window than Transformer encoders, they present an interesting opportunity for moral foundation prediction in long, unlabeled documents. Future research that explores this approach should however be careful to account for ethical implications and potential biases of these models.

BIBLIOGRAPHY

Alva Principe, Renzo, Nicola Chiarini, and Marco Viviani (2025). "Long Document Classification in the Transformer Era: A Survey on Challenges, Advances, and Open Issues." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 15.2, e70019.

Araque, Oscar, Lorenzo Gatti, and Kyriaki Kalimeri (2020). "MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction." In: *Knowledge-based systems* 191, p. 105184.

Atari, Mohammad, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani (2023). "Morality beyond the WEIRD: How the nomological network of morality varies across cultures." In: *Journal of Personality and Social Psychology*.

Beiró, Mariano Gastón, Jacopo D'Ignazi, Victoria Perez Bustos, María Florencia Prado, and Kyriaki Kalimeri (2023). "Moral narratives around the vaccination debate on facebook." In: *Proceedings of the ACM Web Conference 2023*, pp. 4134–4141.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer." In: *arXiv:2004.05150*.

Bottoms, Anthony and Justice Tankebe (2012). "Beyond procedural justice: A dialogic approach to legitimacy in criminal justice." In: *J. Crim. l. & Criminology* 102, p. 119.

Bucher, Martin Juan José and Marco Martini (2024). "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification." In: *arXiv preprint arXiv:2406.08660*.

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

Burlone, Nathalie and Rebecca Grace Richmond (2018). "Between morality and rationality: Framing end-of-life care policy through narratives." In: *Policy Sciences* 51, pp. 313–334.

Cannon, Peter Robert, Simone Schnall, and Mathew White (2011). "Transgressions and expressions: Affective facial muscle activity predicts moral judgments." In: *Social psychological and personality science* 2.3, pp. 325–331.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos (Nov. 2020). "LEGAL-BERT: The Muppets straight out of Law School." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2898–2904. DOI: 10.18653/v1/2020.findings-emnlp.261. URL: https://aclanthology.org/2020.findings-emnlp.261/.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning (2020). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." In: *International Conference on Learning Representations*.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks." In: *Machine learning* 20, pp. 273–297.

Dai, Xiang, Ilias Chalkidis, Sune Darkner, and Desmond Elliott (Dec. 2022). "Revisiting Transformer-based Models for Long Document Classification." In: *Findings of the Association for Computa-*

tional Linguistics: EMNLP 2022. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7212–7230. DOI: 10.18653/v1/2022.findings-emnlp.534. URL: https://aclanthology.org/2022.findings-emnlp.534/.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber (2017). "Automated hate speech detection and the problem of offensive language." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1, pp. 512–515.

Dehghani, Morteza, Kenji Sagae, Sonya Sachdeva, and Jonathan Gratch (2014). "Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque"." In: *Journal of Information Technology & Politics* 11.1, pp. 1–14.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Fang, Hongchao, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie (2020). "Cert: Contrastive self-supervised learning for language understanding." In: *arXiv preprint arXiv:2005.12766*.

Fiske, Alan Page (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. Free Press.

Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, p. 378.

Frimer, Jeremy A, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehgani (2019). "Moral foundations dictionary for linguistic analyses 2.0." In: *Unpublished manuscript*. Available at https://doi.org/10.17605/OSF.IO/EZN37.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky (2016). "Domain-adversarial training of neural networks." In: *Journal of machine learning research* 17.59, pp. 1–35.

Garbas, Lukas, Max Ploner, and Alan Akbik (2024). "TransformerRanker: A Tool for Efficiently Finding the Best-Suited Language Models for Downstream Classification Tasks." In: *CoRR* abs/2409.05997. URL: https://doi.org/10.48550/arXiv.2409.05997.

Graham, Jesse (2010). "Left gut, right gut: Ideology and automatic moral reactions." PhD thesis. University of Virginia Charlottesville.

Graham, Jesse and Jonathan Haidt (2012). "Sacred values and evil adversaries: A moral foundations approach." In.

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

Graham, Jesse, Jonathan Haidt, and Brian A Nosek (2009). "Liberals and conservatives rely on different sets of moral foundations." In: *Journal of personality and social psychology* 96.5, p. 1029.

Graham, Jesse, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto (2011). "Mapping the moral domain." In: *Journal of personality and social psychology* 101.2, p. 366.

Grosfeld, Eva, Daan Scheepers, and Armin Cuyvers (2024). "Mapping the moral foundations of the European Union: Why a lack of moral diversity may undermine perceived EU legitimacy." In: *PNAS nexus* 3.8, p. 282.

Guo, Siyi, Negar Mokhberian, and Kristina Lerman (2023). "A data fusion framework for multi-domain morality learning." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 17, pp. 281–291.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith (July 2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740. URL: https://aclanthology.org/2020.acl-main.740/.

Haidt, Jonathan (2001). "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." In: *Psychological review* 108.4, p. 814.

Haidt, Jonathan (2012). "The righteous mind: Why good people are divided by politics and religion." In: *New York Pantheon*.

Haidt, Jonathan and Jesse Graham (2007). "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize." In: *Social justice research* 20.1, pp. 98–116.

Haidt, Jonathan and Craig Joseph (2004). "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues." In: *Daedalus* 133.4, pp. 55–66.

Harper, Craig A and Todd E Hogue (2019). "The role of intuitive moral foundations in Britain's vote on EU membership." In: *Journal of Community & Applied Social Psychology* 29.2, pp. 90–103.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd (2020). "spaCy: Industrial-strength Natural Language Processing in Python." In: DOI: 10.5281/zenodo.1212303.

Hoover, Joe, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. (2020). "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment." In: *Social Psychological and Personality Science* 11.8, pp. 1057–1071.

Hopp, Frederic R, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber (2021). "The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text." In: *Behavior research methods* 53, pp. 232–246.

Hu, Jingzhao, Hao Zhang, Yang Liu, Richard Sutcliffe, and Jun Feng (2022). "BBW: a batch balance wrapper for training deep neural networks on extremely imbalanced datasets with few minority samples." In: *Applied Intelligence*, pp. 1–16.

Hutto, Clayton and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1, pp. 216–225.

Jackson, Jonathan, Ben Bradford, Mike Hough, Andy Myhill, Paul Quinton, and Tom R Tyler (2012). "Why do people comply with the law? Legitimacy and the influence of legal institutions." In: *British journal of criminology* 52.6, pp. 1051–1071.

Jackson, Jonathan, Mike Hough, Ben Bradford, and Jouni Kuha (2015). "Empirical legitimacy as two connected psychological states." In: *Trust and legitimacy in criminal justice: European perspectives*, pp. 137–160.

Joachims, Thorsten (1998). "Text categorization with support vector machines: Learning with many relevant features." In: *European conference on machine learning*. Springer, pp. 137–142.

Johnson, Kristen and Dan Goldwasser (2018). "Classification of moral foundations in microblog political discourse." In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 720–730.

Kahneman, Daniel (2011). "Thinking, Fast and Slow/Farrar." In: *Straus and Giroux*.

Kobbe, Jonathan, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt (2020). "Exploring morality in argumentation." In: Association for Computational Linguistics, ACL.

Krippendorff, Klaus (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data." In: *biometrics*, pp. 159–174.

Lewis, Paul G (2019). "Moral foundations in the 2015-16 US presidential primary debates: The positive and negative moral vocabulary of partisan elites." In: *Social Sciences* 8.8, p. 233.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár (2017). "Focal loss for dense object detection." In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liscio, Enrico, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukannaiah (2022). "Cross-Domain Classification of Moral Values." In: *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2727–2745.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach." In: *arXiv preprint arXiv:1907.11692*.

McAdams, Dan P, Michelle Albaugh, Emily Farber, Jennifer Daniels, Regina L Logan, and Brad Olson (2008). "Family metaphors and moral intuitions: how conservatives and liberals narrate their lives." In: *Journal of personality and social psychology* 95.4, p. 978.

Mokhberian, Negar, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman (2022). "Noise audits improve moral foundation classification." In: *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, pp. 147–154.

Mucciaroni, Gary (2011). "Are debates about "morality policy" really about morality? Framing opposition to gay and lesbian rights." In: *Policy Studies Journal* 39.2, pp. 187–216.

Nguyen, Tuan Dung, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie (2024). "Measuring moral dimensions in social media with mformer." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 18, pp. 1134–1147.

Pappagari, Raghavendra, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak (2019). "Hierarchical transformers for long document classification." In: *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. ieee, pp. 838–844.

Preniqi, Vjosa, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri (2024). "Moralbert: A fine-tuned language model for capturing moral values in social discussions." In: *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pp. 433–442.

Roy, Shamik and Dan Goldwasser (2021). "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*, pp. 1–13.

Schwartz, Shalom H and Wolfgang Bilsky (1990). "Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications." In: *Journal of personality and social psychology* 58.5, p. 878.

Seppälä, Selja (Aug. 2019). *EU Regulation Corpus Compiler*. Version v1.0.0. DOI: 10.5281/zenodo.4592251. URL: https://doi.org/10.5281/zenodo.4592251.

Shweder, RA (1987). *Culture and moral development*.

Sim, Julius and Chris C Wright (2005). "The kappa statistic in reliability studies: use, interpretation, and sample size requirements." In: *Physical therapy* 85.3, pp. 257–268.

Sommerer, Thomas and Hans Agné (2018). "Consequences of legitimacy in global governance." In: *Legitimacy in global governance: Sources, processes, and consequences*, pp. 153–168.

Takikawa, Hiroki and Takuto Sakamoto (2017). "Moral Foundations of Political Discourse: Comparative Analysis of the Speech Records of the US Congress and the Japanese Diet." In: *International Conference on Computational Social Science IC*. Vol. 2, p. 2.

Tan, Enhao and Haowei Liu (2022). "Performance Comparison of Seven Pretrained Models on a text classification task." In: *Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning*, pp. 8–12.

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

Treib, Oliver (2021). "Euroscepticism is here to stay: what cleavage theory can teach us about the 2019 European Parliament elections." In: *Journal of European public policy* 28.2, pp. 174–189.

Tsirmpas, Dimitrios, Ioannis Gkionis, Georgios Th Papadopoulos, and Ioannis Mademlis (2024). "Neural natural language processing for long texts: A survey on classification and summarization." In: *Engineering Applications of Artificial Intelligence* 133, p. 108231.

Tyler, Tom and Jonathan Jackson (2013). "Future challenges in the study of legitimacy and criminal justice." In: *Yale Law School, Public Law Working Paper* 264.

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.

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

Wachsmuth, Henning, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein (2017). "Computational argumentation quality assessment in natural language." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 176–187.

Weber, René, J Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini (2021). "Extracting latent moral information from text narratives: Relevance, challenges, and solutions." In: *Computational Methods for Communication Science*. Routledge, pp. 39–59.

Wendell, Dane G and Raymond Tatalovich (2021). "Classifying public policies with moral foundations theory." In: *Policy Sciences* 54, pp. 155–182.

Werts, Han, Peer Scheepers, and Marcel Lubbers (2013). "Euro-scepticism and radical right-wing voting in Europe, 2002–2008: Social cleavages, socio-political attitudes and contextual characteristics determining voting for the radical right." In: *European Union Politics* 14.2, pp. 183–205.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. (2020). "Transform-

ers: State-of-the-art natural language processing." In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.

Zangari, Lorenzo, Candida M Greco, Davide Picca, and Andrea Tagarelli (2025). "ME2-BERT: Are Events and Emotions what you need for Moral Foundation Prediction?" In: *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9516–9532.

Zhang, Yuan, Regina Barzilay, and Tommi Jaakkola (Dec. 2017). "Aspect-augmented Adversarial Networks for Domain Adaptation." In: *Transactions of the Association for Computational Linguistics* 5, pp. 515–528. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00077. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00077/1567529/tacl_a_00077.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00077.

APPENDIX

A EU DOCUMENT ANNOTATION INSTRUCTIONS

In the following, we provide the full instructions that were shown to annotators during the annotation process of the EU documents.

OBJECTIVE: Annotate each EU law document with the moral foundations that are expressed by the document.

MORAL FOUNDATIONS THEORY (MFT) suggests that human morality is rooted in several innate psychological systems, which have developed in humans because of evolutionary advantages, mostly in creating a functioning society. You can find more information about MFT at `https://moralfoundations.org/`.

For this annotation task, we are interested in the original five moral foundations, which always come in pairs including a virtue and a vice:

- **Care:** Concern for the physical and emotional well-being of others, particularly vulnerable individuals.
  Examples: Crying when seeing others suffering; Helping people in an accident.
- **Harm:** Actions or intentions to cause physical or psychological damage to people or animals.
  Examples: Hitting someone; Not stopping to help an injured person.
- **Fairness:** Distribution of resources, opportunities, and responsibilities in an equal or proportional way as well as punishing wrongdoers in a way that is accepted by society.
  Examples: Ensuring equal success opportunities; Demanding a severe penalty for a murderer.
- **Cheating:** Unequal or unproportional distribution of resources and opportunities; dishonesty, favoritism and unfair punishments (too severe/mild).
  Examples: Breaking a promise; Two people earning a different salary for the same job.
- **Loyalty:** Allegiance to a group (e.g. nation, team, race) even at cost to the self as well as being biased against outgroups.
  Examples: Wearing clothes with the logo of your team; Going to war for your country.

- **Betrayal:** Actions or intentions to violate group expectations or to sacrifice group interests in favor of the individual.
  Examples: Selling company secrets to a competitor; Leaving a sports team.
- **Authority:** Respect for hierarchical structures, legitimate leadership, social norms and traditional customs.
  Examples: Following the orders of leaders; Covering your head when entering a religious building.
- **Subversion:** Attempts to undermine or challenge traditions, legitimate leaders or the power of any group the person belongs to.
  Examples: Protesting against the government; Opposing arranged marriage.
- **Sanctity:** Beliefs about purity, divinity, and the sacredness of certain entities.
  Examples: Going to church every Sunday; Caring for your hygiene.
- **Degradation:** Actions or objects considered impure, contaminated or disgusting.
  Examples: Burning bibles; General feeling of disgust towards illness or excrements.

ANNOTATION PROCESS

- Please read the shown document and select the moral foundations on the right, if any of them are expressed by the text. You can also use the document's metadata on the left as context.
- After selecting the respective labels, please click the 'Save Annotation' button.
- If you do not want to label a certain document, click the 'Skip' button. Nothing will be saved in this case. This is different from saving without selecting any labels. Please also use this option when the document is broken in some way.
- The first 10 documents are the same for every annotator and cannot be skipped. After that, you will only see new unlabeled documents. Please complete at least 20 documents, but of course doing more is much appreciated!

ANNOTATION HELP

To make the annotation work a bit easier, we have used a dictionary-based method to highlight some words in the text that can be an indicator for certain moral foundations.

Please note that the meaning of a single word depends on the context it's used in, which is not accounted for in the highlighting process. So highlighted words may also not have any moral meaning at all and especially EU documents can use words outside of their everyday context. For example there are documents discussing the trade of 'oilseed rape', which is not a moral topic at all, but the dictionary detects 'rape' as an extremely moral term.

ADDITIONAL NOTES

- Since the documents can be repetitive or long-winded, it is totally okay to skim over parts of the text, as long as you understand the general content.
- According to Moral Foundations Theory we make moral judgements intuitively rather than by thinking rationally. Your annotation decisions can follow the same principle. However, please only use information provided in the text and as little context knowledge as possible.
- Since EU documents tend to be extremely long, we cut off the documents after a certain length. While the shown texts should never end in the middle of a sentence, please be aware that you usually do not see the whole document.
- Each document can express multiple moral foundations, even contrary ones like care and harm. Especially long documents can cover a lot of different aspects.
- A document can also express no moral foundation at all. In that case please click the save button without selecting any labels.
- You can hide these instructions by clicking on the "Instructions" title.

DATA PERMISSIONS By taking part in this annotation task, you consent to your responses being used for academic purposes. You also agree that your responses may be shared publicly for academic use, with all personally identifiable information removed.

B IMPACT OF THE 50% AGREEMENT RULE

	In-domain		Cross-domain			
	MFTC & MFRC		MFNC–paragraphs		MFNC–documents	
	Rule not applied	Rule applied	Rule not applied	Rule applied	Rule not applied	Rule applied
Care	**0.560 ± 0.027**	0.398 ± 0.040	**0.246 ± 0.050**	0.129 ± 0.014	**0.718 ± 0.026**	0.437 ± 0.138
Harm	**0.610 ± 0.010**	0.410 ± 0.033	**0.346 ± 0.114**	0.311 ± 0.045	**0.728 ± 0.022**	0.602 ± 0.046
Fairness	0.472 ± 0.017	**0.500 ± 0.042**	**0.232 ± 0.056**	0.163 ± 0.051	**0.747 ± 0.039**	0.544 ± 0.128
Cheating	**0.552 ± 0.016**	0.422 ± 0.016	**0.345 ± 0.074**	0.211 ± 0.046	**0.776 ± 0.015**	0.540 ± 0.043
Loyalty	0.505 ± 0.017	**0.604 ± 0.023**	**0.194 ± 0.028**	0.028 ± 0.014	**0.754 ± 0.039**	0.200 ± 0.064
Betrayal	**0.355 ± 0.010**	0.344 ± 0.149	**0.219 ± 0.052**	0.077 ± 0.070	**0.715 ± 0.057**	0.273 ± 0.277
Authority	**0.401 ± 0.006**	0.238 ± 0.117	**0.202 ± 0.041**	0.058 ± 0.031	**0.743 ± 0.028**	0.289 ± 0.111
Subversion	**0.461 ± 0.020**	0.446 ± 0.031	**0.287 ± 0.096**	0.041 ± 0.010	**0.767 ± 0.073**	0.181 ± 0.050
Purity	**0.415 ± 0.023**	0.158 ± 0.078	**0.163 ± 0.042**	0.000 ± 0.000	**0.602 ± 0.114**	0.000 ± 0.000
Degradation	**0.375 ± 0.021**	0.362 ± 0.229	**0.201 ± 0.041**	0.002 ± 0.004	**0.654 ± 0.087**	0.007 ± 0.016
Macro Avg.	**0.471 ± 0.004**	0.388 ± 0.038	**0.243 ± 0.020**	0.102 ± 0.008	**0.720 ± 0.014**	0.307 ± 0.038

Table 15: We explore the performance impact of restricting the MFTC and MFRC to only the labels with at least 50% agreement between annotators. To compare the two settings of using this rule and not using it, we employ the tuned architecture (i.e. optimal hyperparameters, removed reconstruction module) with BERT as feature extractor and the label aggregation strategy. We report the averaged F1 score on the validation split of the respective datasets over five repetitions. It can be seen that applying the 50% agreement rule substantially reduces the performance of the model.

C MFNC–DOCUMENTS AS TARGET DOMAIN FOR LONGFORMER

	Cross-domain dataset	
	MFNC–documents	MFNC–paragraphs
Care	0.040 ± 0.048	$\mathbf{0.164 \pm 0.047}$
Harm	0.231 ± 0.136	$\mathbf{0.354 \pm 0.119}$
Fairness	0.044 ± 0.046	$\mathbf{0.224 \pm 0.032}$
Cheating	0.079 ± 0.052	$\mathbf{0.185 \pm 0.049}$
Loyalty	0.046 ± 0.055	$\mathbf{0.205 \pm 0.115}$
Betrayal	0.000 ± 0.000	$\mathbf{0.061 \pm 0.024}$
Authority	0.006 ± 0.013	$\mathbf{0.232 \pm 0.104}$
Subversion	0.104 ± 0.043	$\mathbf{0.323 \pm 0.039}$
Purity	0.000 ± 0.000	$\mathbf{0.007 \pm 0.015}$
Degradation	0.007 ± 0.016	$\mathbf{0.021 \pm 0.032}$
Macro Avg.	0.056 ± 0.015	$\mathbf{0.178 \pm 0.021}$

Table 16: We explore whether the Longformer approach should use the MFNC–documents or the MFNC–paragraphs as cross-domain data during training. To compare the two settings, we employ the same setup as in Section 5.5. We report the averaged F1 score on the validation split of the respective datasets over five repetitions. Clearly, using the MFNC–paragraphs during training results in a better performance.

D REPETITIONS OF DOMAIN TRANSFER SUCCESS ANALYSIS

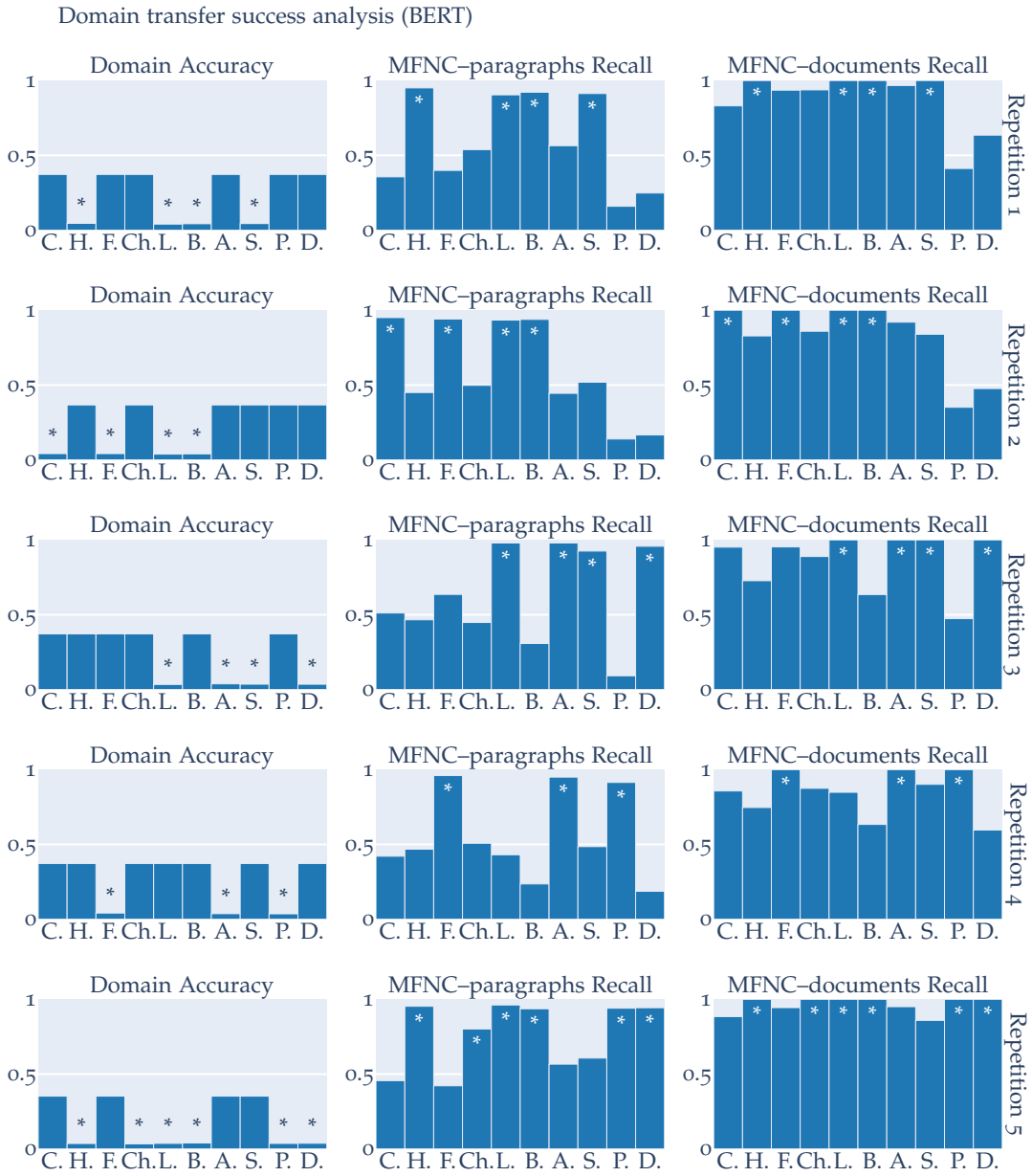

Figure 9: Here we provide the domain success analysis from Section 5.7 for all five conducted repetitions where BERT was used as the feature extractor. The observations we discussed in the main text are consistent across all repetitions.

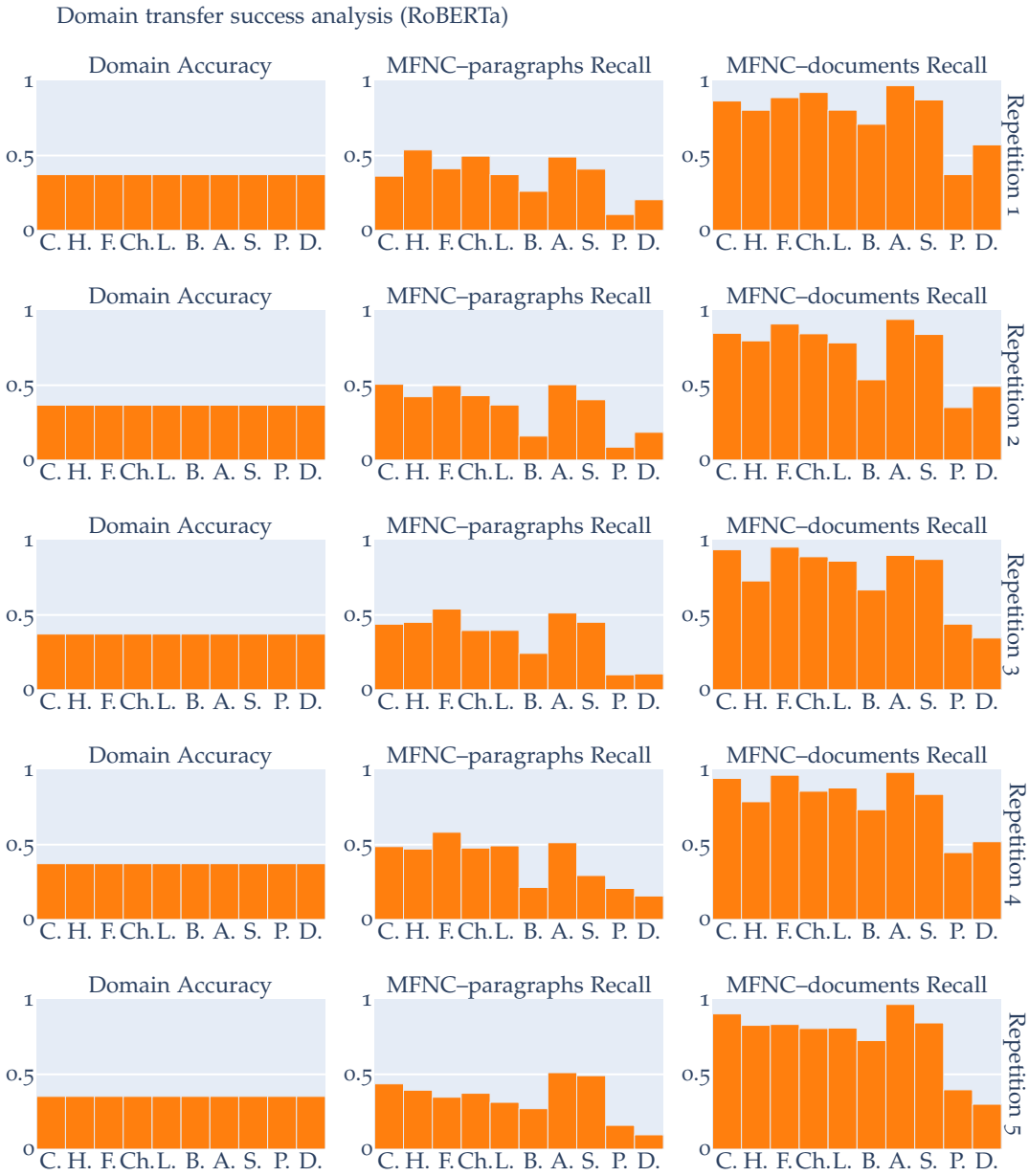

Figure 10: Here we provide the domain success analysis from Section 5.7 for all five conducted repetitions where RoBERTa was used as the feature extractor. Over all repetitions, we do not observe any failed domain transfers.