Advancing CSR Theme and Topic Classification: LLMs and Training Enhancement Insights

Jens Van Nooten[*], **Andriy Kosar**[**|*], **Guy De Pauw**[**], **Walter Daelemans**[*]
[*]University of Antwerp (CLiPS), [**]Textgain
Antwerp, Belgium
{jens.vannooten, walter.daelemans}@uantwerpen.be
{andrew, guy}@textgain.com

Abstract

In this paper, we present our results of the classification of Corporate Social Responsibility (CSR) Themes and Topics shared task, which encompasses cross-lingual multi-class classification and monolingual multi-label classification. We examine the performance of multiple machine learning (ML) models, ranging from classical models to pre-trained large language models (LLMs), and assess the effectiveness of Data Augmentation (DA), Data Translation (DT), and Contrastive Learning (CL). We find that state-of-the-art generative LLMs in a zero-shot setup still fall behind on more complex classification tasks compared to fine-tuning local models with enhanced datasets and additional training objectives. Our work provides a wide array of comparisons and highlights the relevance of utilizing smaller language models for more complex classification tasks.

Keywords: multi-class classification, multi-label classification, cross-lingual classification, CSR

1. Introduction

The landscape of Corporate Social Responsibility (CSR) is increasingly becoming a pivotal aspect of how businesses operate and are perceived in the global market (Wen and Deltas, 2022). Significant regulations have been instrumental in shaping the CSR framework. For a comprehensive history of CSR regulation, consult Wen and Deltas (2022).

These regulations have increased the liability of companies regarding sustainability non-compliance, making it imperative for them to not only be aware of but also manage and anticipate such issues effectively. However, even with mandatory or voluntary reporting, not all pertinent information is disclosed or reported and consequently leveraged for company evaluation due to CSR-related information being scattered across different media sources, languages and formats. This leads to challenges in its identification and analysis. As a result, there is a critical need for efficient methods to detect and classify this diverse information in order to reinforce corporate compliance and enhance stakeholder decision-making.

In response to this growing need and interest in processing and analyzing CSR content, our study addresses the complexities of detecting and classifying CSR content through participation in the "Cross-lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics" shared task (Nayekoo et al., 2024). The task facilitates cross-lingual CSR theme detection and fine-grained topic classification, specifically targeting the Environment (ENV) and Labour and Human Rights (LAB) themes across English, French, and simplified Chinese. The theme classification is approached as a multi-class problem, and the topic classification within these themes is framed as a multilabel classification task. Our evaluation extends to various text representations and ML models, encompassing both traditional approaches and Large Language Models (LLMs), utilizing pre-trained models for ZS classification and Fine-tuning. Additionally, we explore the potential of enhancement techniques like Data Augmentation (DA) and Contrastive Learning (CL) to improve performance.

In the following sections, we delve into the methodology employed in our study, the experimental setup, the results and analysis of our findings, and the implications of our research for the field of CSR content processing and classification.

2. Previous Work

Text Classification The field of text classification, encompassing both multi-class and multi-label types, has experienced significant evolution over the past decade. This evolution has been particularly notable in three key areas: model types, text representation, and training methods. The advent of LLMs, starting with BERT, has transformed the landscape by introducing advanced model architectures, enhancing text representation through context-aware embeddings, and pioneering efficient training methodologies that leverage pre-trained models for fine-tuning or even enable zero-shot learning capabilities. For an overview of the diverse approaches and developments in multi-

These authors contributed equally to this work.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 292–305
20 May, 2024.

class and multi-label text classification, we refer to the comprehensive surveys conducted by Li et al. (2022), Gasparetto et al. (2022), Chen et al. (2022) and Bogatinovski et al. (2022), which cover both existing approaches and the latest advancements.

NLP for CSR In addition to the broad advancements in text classification, there has not been much research conducted on applying these techniques to the Corporate Social Responsibility (CSR) domain, with a few exceptions. Most of the work was conducted for the automatic analysis of Corporate Sustainability Reports: Shahi et al. (2011, 2014) applied multi-label text classification to classify reports according to the Global Reporting Initiative Index. Castellanos et al. (2015) applied neural networks, decision trees, and a memory-based learning algorithm, to classify parts of the report according to the five dimensions of the Sustainability Accounting Standards Board.

CSR has recently attracted more attention in the context of NLP, exemplified by the First Computing Social Responsibility Workshop (CSR-NLP I 2022) (Wan and Huang, 2022). However, to the best of our knowledge, there has been limited progress in classifying publicly accessible information on the internet across diverse textual genres, including but not limited to news articles, company briefs/newsletters, and industry reports.

LLMs for Text Classification LLMs have been used widely in the field of text classification ever since the advent of BERT. In the following years, countless new models have been released that yield state-of-the-art results on a multitude of classification and generation tasks, such as LLAMA (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Gemini (Team et al., 2023) and GPT (Radford et al., 2018). Recently, GPT-3.5 (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023) have sparked the interest of many researchers, leading to a great deal of work being dedicated to their applications for classification, besides generation. For a more comprehensive overview, consult Minaee et al. (2024). As evidenced in Peskine et al. (2023) and De Langhe et al. (2024), prompting generative models for more complex classification tasks such as multi-label classification can be quite challenging, leading to inferior performance compared to fine-tuned encoders.

Data Augmentation To address the issue of data imbalance and data scarcity, multiple data augmentation techniques have been leveraged, including generating synthetic data with state-of-the-art GPT models (Van Nooten and Daelemans, 2023; Sufi, 2024; Kumar et al., 2020; Zhang et al., 2020), reaching superior performance compared to other data augmentation methods.

Contrastive Learning Contrastive Learning aims to maximize the distance between dissimilar texts and minimize the distance between similar pairs in the embeddings space. Some studies explore contrastive losses (CL) for multi-class classification (Pan et al., 2022) and multi-label classification (U et al., 2023; Lin et al., 2023) using variants of Supervised Contrastive Loss (SCL) (Khosla et al., 2021) or NT-XENT (Sohn, 2016).

In this work, we aim to provide a wide range of baselines for multi-class and multi-label CSR text classification. We hypothesise that fine-tuning smaller language models can outperform more recent generative LLMs for more complex classification tasks. Moreover, we also hypothesise that CL can further improve performance and that generative LLMs produce useful synthetic data to further enhance performance of classification models, as previous work indicates.

3. Datasets

Shared Task The shared task is divided into two subtasks: cross-lingual, multi-class classification for CSR theme recognition (one dataset) and monolingual multi-label text classification (two datasets) of CSR topics for Environment (ENV) and Labour and Human Rights (LAB) themes. These datasets comprise lists of URLs for English texts, each associated with relevant labels. Table 1 provides statistics for each dataset.

Data Collection and Cleaning The texts in the training dataset were scraped using the Trafilatura library (Barbaresi, 2021). URLs that could not be successfully scraped were excluded from the training dataset. Given that a significant portion of the data contained artifacts potentially detrimental to the training of the models (such as URLs, external links or other irrelevant text), we employed GPT-3.5[1] for data cleaning. The resulting cleaned texts were checked manually. After cleaning the data and removing duplicates, 675 of 699 texts remained. The specific prompt that was used is described in Appendix B.1. The test data were scraped using the Boilerpy library[2], following the organizers' recommendation.

4. Methododogy

Classification Models In our study, we evaluated the performance of a wide range of classifi-

[1] *gpt-3.5-turbo-0613*, `https://platform.openai.com/docs/models/gpt-3-5-turbo`

[2] `https://pypi.org/project/boilerpy3/`

Dataset	type	n classes	labels per text	n train	n test
Themes	Multi-class	4	1	1,515	618
ENV	multi-label	9	1.53	675	157
LAB	multi-label	9	1.35	500	149

Table 1: Datasets' statistics.

cation models, which differ significantly in terms of model complexity, to identify those most suitable for tasks with limited training data. As a baseline, we chose the SVM model combined with TF-IDF and OpenAI text embeddings. Additionally, we included Zero-shot (ZS) text classification with GPT-3.5 and GPT-4 models (Radford et al., 2018) as baselines. Examples of prompts for ZS text classification are provided in Appendix B.3 and B.4.

We further expand our model repertoire by incorporating models with more complex architectures, specifically, the Multi-Layer Perceptron (MLP) and an LSTM, while maintaining the same text representation strategies to ensure a consistent comparison basis. Detailed information on the optimal hyperparameters identified for these models can be found in Appendix D.

For the themes dataset, we utilized multilingual pre-trained language models, such as Multi-Lingual DistilBERT[3] (Sanh et al., 2020), XLM-RoBERTa, and XLM-RoBERTa-large[4] (Conneau et al., 2020). For the multi-label datasets, we employed DistilBERT, BERT (Devlin et al., 2019), RoBERTa, and RoBERTa-large (Liu et al., 2019). All models were trained with a batch size of 8 and 2 gradient accumulation steps. Where appropriate, we repeated each experiment with three random seeds. The optimal hyperparameters for all models are detailed in Appendix D.

Data Paraphrasing and Translation The training data for each task was expanded using various methods. Every entry in the datasets was paraphrased using Mixtral[5], effectively doubling the size of the training data. A detailed description of the prompt used can be found in Appendix B.2. Additionally, for the cross-lingual multi-class task, we opted for translating the data to French and simplified Chinese using the Google Translate API[6]. These languages were selected because the model was to be tested on them. By translating the data into these two additional languages, the size of the training data was tripled. The synthetic data was incorporated into the training dataset.

Contrastive Learning We train additional models with a contrastive loss. For the multi-label variant, we follow Lin et al. (2023) and select positive and negative in-batch samples for a given anchor by calculating the Jaccard Index (JI) between binary label vector pairs. If JI is greater than the threshold hyperparameter, the sample is considered a positive. To allow CL to work with relatively small batches, all possible pair combinations are constructed in a batch to maximize the information gained from the contrastive loss.

We devise a variant of NT-Xent (Sohn, 2016) that allows for multiple positives to be taken into account per batch. In essence, we calculate the Binary Cross-Entropy loss between two vectors with length n, where n is the number of possible combinations in a batch: vector α, which is a binary vector that denotes whether a pair is positive or negative, and vector β, which is a vector that contains the cosine distance between the two in-batch samples in a pair (cf. Eq. 1 and 2 in Appendix C). The goal is to minimize the cosine distance between samples in positive pairs and maximize the distance between samples in negative pairs, which leads to a decrease in BCE. The resulting loss is then weighted and added to the classification loss.

Evaluation All models are evaluated in a five-fold stratified cross-validation setup. To stratify the multi-label splits, we employ the strategy described in Sechidis et al. (2011). The corresponding paraphrases and translations of a certain training fold were added during training so no indirect data leakage would occur. As evaluation metrics, micro-averaged and macro-averaged F1 are used.

5. Results

Subtask A The cross-validation results for Subtask A are summarised in Table 2[7]. It can be observed that smaller models with less complex training methods, such as keyword-based learning with SVM, already achieve respectable results[8], though the larger models and more complex models generally achieve the best results. Both CL and adding translations to the training data generally yields improvements in terms of macro- and micro-averaged F1 performance. However, data translation alone yields the best results, which is especially beneficial for learning the minority class "SUP".

Interestingly, we observe that the ZS experiments with GPT-3.5 and GPT-4 yield inferior results, thus indicating that the fine-tuned models benefit from learning the text-specific features during training.

[3] *distilbert/distilbert-base-multilingual-cased*

[4] *FacebookAI/xlm-roberta-(base/large)*

[5] *mistralai/Mixtral-8x7B-Instruct-v0.1*, https://docs.together.ai/docs/inference-models

[6] https://cloud.google.com/translate/docs/basic/translating-text

[7] Confusion matrices for all models and datasets can be found in Appendix F.

[8] These models are solely trained on English data and are not to be deployed on the multilingual test set.

Th.	tf-idf + SVM	ada-003 + SVM	ada-003 + MLP	ada-003 + LSTM	GPT-3	GPT-4	DB	DB + CL	DB + DA	DB + CL + DA	BERT	BERT + CL	BERT + DA	BERT + CL + DA	RB-Lg	RB-Lg + CL	RB-Lg + DA	RB-Lg + CL + DA
ENV	94.59 (± 0.88)	94.59 (± 0.88)	95.93 (± 0.71)	96.07 (± 0.58)	92.44 (± 0.79)	93.69 (± 1.6)	94.4 (± 1.4)	94.54 (± 1.62)	95.42 (± 0.98)	95.73 (± 0.89)	95.34 (± 1.39)	95.62 (± 1.33)	96.45 (± 0.96)	96.46 (± 0.89)	95.58 (± 1.1)	95.28 (± 2.65)	97.01 (± 0.63)	97.06 (± 0.77)
FBP	85.17 (± 4.01)	85.17 (± 4.01)	91.02 (± 3.86)	91.88 (± 2.64)	74.76 (± 3.78)	81.52 (± 5.18)	79.04 (± 6.63)	77.01 (± 7.24)	86.8 (± 3.62)	87.61 (± 2.39)	87.07 (± 4.97)	89.07 (± 5.97)	91.43 (± 3.58)	92.03 (± 3.28)	84.0 (± 31.26)	82.9 (± 36.58)	94.25 (± 2.1)	94.37 (± 1.98)
LAB	94.42 (± 1.66)	94.42 (± 1.66)	96.12 (± 0.88)	96.4 (± 0.72)	90.23 (± 1.5)	94.02 (± 0.98)	93.7 (± 1.45)	93.73 (± 1.48)	94.58 (± 0.76)	95.05 (± 0.8)	95.77 (± 0.69)	96.08 (± 0.73)	96.43 (± 1.09)	96.28 (± 1.26)	94.94 (± 3.85)	95.25 (± 3.66)	97.15 (± 0.88)	96.98 (± 0.71)
SUP	52.5 (± 23.09)	52.5 (± 23.09)	56.92 (± 18.8)	67.06 (± 17.62)	42.1 (± 23.49)	47.72 (± 15.69)	0.0 (± 0.0)	0.0 (± 0.0)	59.89 (± 24.61)	58.29 (± 15.72)	0.0 (± 0.0)	0.0 (± 0.0)	67.0 (± 13.42)	61.54 (± 10.27)	0.0 (± 0.0)	2.0 (± 7.13)	72.54 (± 13.49)	70.71 (± 15.06)
mic	92.81 (± 1.28)	92.81 (± 1.28)	94.72 (± 0.93)	95.16 (± 0.55)	88.09 (± 0.93)	90.96 (± 1.77)	91.27 (± 1.45)	91.11 (± 1.55)	93.36 (± 0.96)	93.64 (± 0.91)	93.47 (± 1.35)	93.97 (± 1.46)	95.18 (± 1.27)	95.05 (± 1.36)	92.96 (± 3.8)	92.87 (± 4.78)	96.22 (± 0.83)	96.13 (± 0.78)
mac	81.67 (± 6.27)	81.67 (± 6.27)	85.0 (± 4.98)	87.85 (± 4.36)	74.88 (± 5.06)	79.24 (± 2.36)	66.79 (± 2.04)	66.32 (± 2.23)	84.17 (± 6.16)	84.17 (± 3.94)	69.55 (± 1.62)	70.19 (± 1.83)	87.83 (± 3.59)	86.58 (± 3.24)	68.63 (± 8.8)	68.86 (± 10.96)	90.24 (± 3.45)	89.78 (± 3.63)
ENV																		
0	47.62 (± 13.88)	47.97 (± 20.94)	49.86 (± 17.46)	50.91 (± 16.13)	16.77 (± 5.14)	23.03 (± 5.35)	52.07 (± 19.2)	51.85 (± 24.23)	57.14 (± 20.91)	56.65 (± 20.04)	53.93 (± 16.06)	52.57 (± 20.2)	61.02 (± 16.66)	56.41 (± 12.48)	57.78 (± 13.44)	65.99 (± 15.79)	60.0 (± 9.82)	62.83 (± 9.43)
1	40.97 (± 11.46)	56.96 (± 7.34)	57.66 (± 3.61)	55.29 (± 7.31)	54.07 (± 7.34)	54.69 (± 6.41)	55.42 (± 8.52)	58.58 (± 10.3)	56.44 (± 13.75)	63.28 (± 11.69)	56.19 (± 11.46)	58.36 (± 8.96)	58.93 (± 12.66)	60.34 (± 9.02)	63.0 (± 10.05)	66.47 (± 9.59)	60.31 (± 10.01)	64.46 (± 11.25)
2	49.39 (± 14.21)	71.01 (± 16.22)	75.77 (± 14.38)	74.27 (± 12.1)	36.84 (± 10.9)	59.41 (± 10.04)	63.79 (± 12.61)	66.0 (± 16.33)	67.09 (± 10.97)	66.23 (± 15.39)	62.63 (± 16.61)	64.95 (± 13.24)	64.67 (± 15.9)	64.88 (± 16.22)	75.57 (± 7.73)	78.47 (± 6.6)	73.93 (± 9.23)	77.78 (± 5.06)
3	79.1 (± 2.62)	80.72 (± 1.34)	81.44 (± 2.24)	82.4 (± 1.63)	80.77 (± 3.39)	84.55 (± 3.84)	83.77 (± 3.51)	83.45 (± 2.95)	82.62 (± 2.98)	83.22 (± 3.12)	83.57 (± 2.38)	83.43 (± 2.35)	82.78 (± 2.65)	82.83 (± 1.98)	83.74 (± 3.59)	85.71 (± 2.41)	84.25 (± 3.36)	84.4 (± 2.43)
4	42.37 (± 4.25)	47.02 (± 6.69)	47.37 (± 6.84)	49.29 (± 6.49)	26.64 (± 6.58)	53.65 (± 3.19)	50.04 (± 6.48)	51.52 (± 4.1)	51.53 (± 4.33)	53.54 (± 2.97)	52.57 (± 6.2)	51.8 (± 4.78)	54.94 (± 4.41)	54.37 (± 4.07)	50.48 (± 6.25)	53.33 (± 3.34)	55.52 (± 5.34)	55.09 (± 5.05)
5	44.92 (± 6.27)	49.42 (± 11.34)	49.02 (± 8.64)	51.32 (± 10.25)	30.68 (± 4.76)	37.84 (± 4.96)	62.59 (± 9.21)	64.14 (± 6.91)	61.43 (± 10.0)	59.17 (± 9.02)	58.22 (± 11.28)	59.44 (± 7.96)	55.51 (± 7.94)	55.81 (± 8.23)	60.76 (± 7.94)	56.2 (± 8.81)	59.1 (± 9.56)	58.77 (± 11.03)
6	62.5 (± 8.63)	60.23 (± 8.87)	62.54 (± 4.75)	65.21 (± 5.52)	18.75 (± 13.37)	54.11 (± 7.32)	66.67 (± 8.74)	67.46 (± 10.21)	68.59 (± 8.41)	66.04 (± 8.96)	64.3 (± 6.79)	65.21 (± 9.37)	63.11 (± 9.51)	62.87 (± 11.18)	69.14 (± 6.99)	71.0 (± 6.7)	70.78 (± 4.94)	68.57 (± 6.92)
7	7.27 (± 8.91)	13.16 (± 12.49)	13.27 (± 6.87)	1.59 (± 5.54)	5.21 (± 6.43)	28.41 (± 4.99)	0.0 (± 0.0)	1.57 (± 5.54)	17.34 (± 9.16)	10.87 (± 11.01)	4.51 (± 7.68)	11.19 (± 12.88)	16.83 (± 9.88)	18.18 (± 9.67)	1.55 (± 4.54)	4.14 (± 7.86)	22.34 (± 13.47)	19.57 (± 10.65)
8	61.07 (± 6.49)	60.44 (± 3.41)	65.36 (± 7.77)	71.17 (± 8.65)	37.02 (± 15.8)	61.51 (± 7.65)	77.57 (± 8.5)	77.75 (± 7.68)	78.01 (± 8.15)	77.37 (± 9.01)	77.09 (± 7.22)	77.66 (± 8.71)	79.08 (± 7.92)	77.44 (± 5.91)	78.76 (± 5.33)	77.42 (± 4.8)	75.66 (± 5.84)	75.86 (± 6.52)
mic	58.53 (± 0.62)	62.22 (± 1.02)	63.55 (± 0.94)	64.96 (± 0.89)	43.6 (± 1.19)	56.12 (± 1.53)	66.55 (± 2.28)	67.16 (± 2.55)	66.29 (± 2.59)	66.84 (± 2.23)	66.28 (± 2.15)	66.51 (± 2.47)	66.24 (± 2.05)	66.24 (± 1.75)	67.85 (± 2.21)	69.19 (± 1.85)	68.66 (± 3.17)	68.85 (± 2.43)
mac	48.36 (± 1.93)	54.1 (± 1.74)	55.81 (± 1.43)	55.71 (± 1.72)	34.08 (± 1.34)	50.8 (± 1.52)	56.88 (± 3.55)	58.04 (± 4.28)	60.02 (± 3.55)	59.6 (± 3.79)	57.0 (± 3.46)	58.29 (± 3.76)	59.65 (± 4.06)	59.24 (± 3.2)	60.09 (± 2.4)	62.08 (± 2.36)	62.43 (± 3.77)	63.04 (± 3.03)
LAB																		
0	22.33 (± 15.79)	39.97 (± 12.38)	42.4 (± 9.98)	46.98 (± 9.51)	26.85 (± 7.38)	21.62 (± 4.13)	47.18 (± 12.43)	43.04 (± 12.28)	46.61 (± 15.42)	48.65 (± 11.72)	54.26 (± 12.79)	52.73 (± 12.25)	52.02 (± 12.63)	57.22 (± 14.71)	57.39 (± 9.02)	62.69 (± 8.95)	58.64 (± 7.22)	59.36 (± 12.0)
1	0.0 (± 0.0)	20.0 (± 40.0)	60.0 (± 32.66)	0.0 (± 0.0)	4.87 (± 0.57)	13.33 (± 26.67)	0.0 (± 0.0)	0.0 (± 0.0)	12.5 (± 24.94)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	21.05 (± 33.99)	30.0 (± 40.0)
2	72.61 (± 2.64)	69.84 (± 5.78)	72.7 (± 5.9)	74.96 (± 5.18)	53.74 (± 9.78)	70.01 (± 3.1)	75.78 (± 6.0)	77.93 (± 4.81)	77.99 (± 3.62)	77.17 (± 3.77)	78.28 (± 3.8)	77.41 (± 3.47)	75.89 (± 6.69)	77.61 (± 5.01)	77.34 (± 2.89)	79.22 (± 2.87)	79.78 (± 4.14)	77.82 (± 3.46)
3	66.66 (± 3.46)	73.57 (± 5.03)	75.6 (± 2.04)	80.36 (± 3.39)	56.72 (± 4.34)	76.46 (± 3.47)	81.75 (± 3.57)	80.72 (± 5.72)	81.32 (± 4.65)	82.04 (± 4.24)	83.02 (± 3.99)	80.71 (± 2.99)	82.48 (± 4.03)	81.55 (± 3.3)	85.76 (± 3.93)	84.83 (± 3.5)	85.63 (± 4.94)	85.04 (± 5.14)
4	0.0 (± 0.0)	13.33 (± 26.67)	68.0 (± 19.04)	0.0 (± 0.0)	16.97 (± 16.49)	47.05 (± 12.6)	23.53 (± 29.48)	33.33 (± 32.66)	55.81 (± 31.89)	66.67 (± 19.12)	6.45 (± 16.63)	6.45 (± 16.63)	42.11 (± 33.26)	46.15 (± 36.24)	23.53 (± 29.48)	0.0 (± 0.0)	28.57 (± 31.43)	42.11 (± 33.26)
5	25.33 (± 17.71)	19.28 (± 11.35)	12.87 (± 11.71)	20.98 (± 15.9)	20.2 (± 2.45)	18.88 (± 1.39)	3.15 (± 6.5)	5.76 (± 9.77)	11.11 (± 11.99)	25.29 (± 12.99)	12.08 (± 12.06)	10.6 (± 10.16)	23.96 (± 14.08)	24.86 (± 16.78)	15.28 (± 14.07)	7.75 (± 10.03)	25.29 (± 15.32)	23.46 (± 13.24)
6	48.28 (± 18.2)	56.03 (± 15.74)	57.74 (± 14.75)	55.71 (± 14.23)	28.81 (± 15.34)	40.89 (± 6.38)	59.68 (± 9.6)	58.87 (± 9.15)	55.97 (± 6.77)	53.94 (± 10.21)	58.68 (± 6.1)	63.24 (± 9.91)	56.3 (± 9.68)	57.03 (± 7.81)	61.21 (± 10.23)	65.81 (± 10.24)	61.41 (± 10.97)	60.0 (± 9.05)
7	38.0 (± 19.39)	10.0 (± 20.0)	10.0 (± 20.0)	0.0 (± 0.0)	31.02 (± 8.76)	24.26 (± 5.95)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	4.0 (± 12.47)	0.0 (± 0.0)	0.0 (± 0.0)	17.54 (± 22.74)	11.32 (± 18.79)	0.0 (± 0.0)	0.0 (± 0.0)	8.0 (± 17.0)	11.54 (± 20.0)
8	60.92 (± 4.07)	74.64 (± 4.94)	76.5 (± 3.78)	80.64 (± 3.5)	57.34 (± 4.02)	64.71 (± 2.46)	74.18 (± 4.89)	72.41 (± 4.34)	71.86 (± 8.45)	71.94 (± 5.21)	75.81 (± 4.72)	76.05 (± 4.11)	75.64 (± 5.62)	76.88 (± 3.77)	81.1 (± 3.32)	82.66 (± 3.83)	79.27 (± 5.29)	78.71 (± 5.42)
mic	57.15 (± 2.82)	63.44 (± 3.47)	65.84 (± 2.87)	69.22 (± 2.34)	36.16 (± 1.76)	47.68 (± 1.58)	67.53 (± 2.93)	66.73 (± 2.03)	66.85 (± 2.65)	67.67 (± 2.48)	69.28 (± 2.45)	68.84 (± 1.89)	68.53 (± 2.96)	69.99 (± 2.69)	72.13 (± 2.16)	73.7 (± 1.62)	72.26 (± 2.13)	71.59 (± 3.59)
mac	37.13 (± 4.8)	41.85 (± 7.22)	52.87 (± 7.49)	39.96 (± 2.26)	32.95 (± 1.94)	41.91 (± 3.57)	40.58 (± 4.35)	41.34 (± 4.52)	45.91 (± 5.67)	47.74 (± 3.86)	40.95 (± 2.47)	40.8 (± 1.97)	47.33 (± 5.76)	48.07 (± 5.13)	44.62 (± 4.83)	42.55 (± 2.01)	49.74 (± 5.57)	52.0 (± 6.61

Table 2: Mean results (F1) and standard deviations across folds and random seeds (if applicable) on the Themes (Th.), ENV and LAB datasets respectively. DB = (multilingual) DistilBERT, RB = (XLM-)RoBERTa. Red = worst score across models, green = best score across models. Consult Appendix E for a label index - label name mapping.

Subtask B The cross-validation results for Subtask B are summarised in Table 2. We observe that the tf-idf approach yields the worst results and that the larger models yield the best results. Additionally, we observe that Contrastive Learning and Data Augmentation generally yield improvements for each base model, indicating that the better separation between class-wise instances in the embedding space is beneficial for learning the task. Moreover, the added paraphrases aid the models especially in predicting uncommon classes. For both of these methods, an increase in true positives, but also false positives is observed across several models. The best macro-averaged results on the ENV dataset are achieved when a combination of the two is used with RoBERTa-large, while the best micro-averaged performance is achieved by training the model with CL.

The LAB dataset is more challenging to classify, as evidenced by the relatively lower scores. We found that RoBERTa-large trained with CL yielded the best micro-averaged performance. However, this model fails to predict some infrequent classes (*Child Labor*, *Ext. Stakeh. Human Rights* and *Soc.l Discr.*), as opposed to the ada-003 + MLP model or models trained with extra data, which each yield a superior macro-averaged performance.

Similar to the results from Subtask A, we observe that the generative models underperform compared to fine-tuned language models on both datasets. This is to be expected, since multi-label classification is challenging with regards to the number of labels that are assigned to a single instance. Such patterns could only be learned by models by including annotation guidelines in the prompt (which we did not have access to) or from the training data itself, which the generative models did not have access to. Fine-tuning generative LLMs on multi-label data could address this issue.

6. Conclusion

In this study, we examined several LLMs for classifying CSR themes and fine-grained CSR topics. We found that even though some smaller, less complex models yield respectable results for both multi-class and multi-label CSR classification, larger fine-tuned models are more successful at performing tasks. ZS experiments with GPT models showed that those models still fall behind on fine-tuned models for multi-label classification. This shortfall can be largely attributed to the complexities of multi-label classification, which demands an understanding of either explicit annotation guidelines or implicit annotator knowledge – insights that are not accessible to LLMs in a ZS context.

7. Acknowledgements

This research was funded by Flanders Innovation & Entrepreneurship (VLAIO), grant HBC.2021.0222 and by the Flemish government under FWO IRI project CLARIAH-VL.

8. Limitations

This study's findings are subject to several limitations. Firstly, the computational cost of running models on a large scale was not considered, which is crucial in practical applications due to resource constraints. Secondly, the choice of prompts for ZS classification and label interpretation may have affected the results, suggesting that exploring different prompting strategies could enhance performance. Thirdly, despite no significant impact observed from testing truncated texts, models capable of processing longer sequences might inherently benefit from more contextual information. These limitations highlight the need for continuous research to refine evaluation methodologies for LLMs, particularly in classifying CSR themes and topics.

9. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. 2022. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.

Arturo Rodríguez Castellanos, Carlos M. Parra, and Monica Chiarini Tremblay. 2015. Corporate social responsibility reports: Understanding topics via text mining. In *Americas Conference on Information Systems*.

Xiaolong Chen, Jieren Cheng, Jingxin Liu, Wenghang Xu, Shuai Hua, Zhu Tang, and Victor S. Sheng. 2022. A survey of multi-label text classification based on deep learning. In *Artificial Intelligence and Security: 8th International Conference, ICAIS 2022, Qinghai, China, July 15–20, 2022, Proceedings, Part I*, page 443–456, Berlin, Heidelberg. Springer-Verlag.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Loic De Langhe, Aaron Maladry, Bram Vanroy, Luna De Bruyne, Pranaydeep Singh, and Els Lefever. 2024. Benchmarking zero-shot text classification for dutch. *Computational Linguistics in the Netherlands Journal*, 13:63–90.

Jacob Devlin, 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. 2022. A survey on text classification algorithms: From text to predictions. *Information*, 13(2).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, 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. *arXiv preprint arXiv:1907.11692*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey.

Yola Nayekoo, Sophia Katrenko, Veronique Hoste, Aaron Maladry, and Els Lefever. 2024. Shared task for cross-lingual classification of corporate social responsibility (csr) themes and topics. In *Proceedings of the Joint Workshop of FinNLP-KDF-ECONLP @ LREC-COLING 2024*, Torino, Italy.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amir Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2014. Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications*, 13:27 pages.

Amir Mohammad Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2011. Analysis of supervised text classification algorithms on corporate sustainability reports. In *Proceedings of 2011 International Conference on Computer Science and Network Technology*, volume 1, pages 96–100.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Fahim Sufi. 2024. Generative pre-trained transformer (gpt) in research: A systematic review on data augmentation. *Information*, 15(2).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Simon Chi Lok U, Jie He, Víctor Gutiérrez-Basulto, and Jeff Z. Pan. 2023. Instances and labels: Hierarchy-aware joint supervised contrastive learning for hierarchical multi-label text classification.

Jens Van Nooten and Walter Daelemans. 2023. Improving Dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada. Association for Computational Linguistics.

Mingyu Wan and Chu-Ren Huang, editors. 2022. *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

Hui Wen and George Deltas. 2022. Global corporate social responsibility reporting regulation. *Contemporary Economic Policy*, 40(1):98–123.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification.

Appendix

A. Class Counts per Dataset

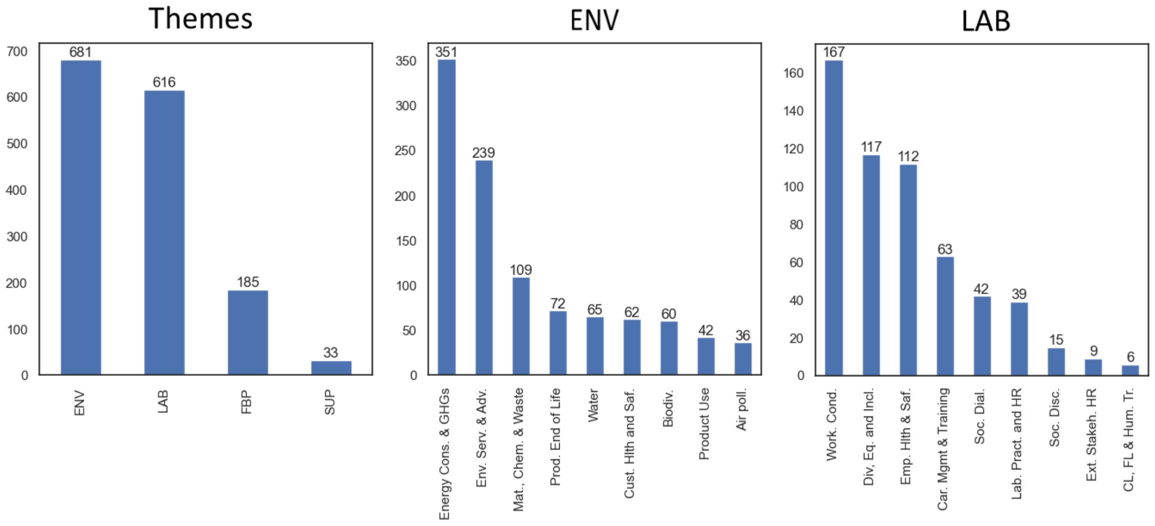

Figure 1: Class counts per (cleaned) dataset.

B. Prompts

B.1. Data Cleaning

"You are a human annotator extracting relevant parts of messy, unstructured texts. Your task is to extract useful parts of texts, like titles, subtitles and paragraphs of texts. The texts will be used for a classification task. It is very important that you ONLY remove the parts of the text that are not useful. The text that you will have to process will be provided in the next message. The output should just be the text, without any other information. Do not generate anything else besides the provided text."

B.2. Data Augmentation

"You are a helpful assistant tasked with creating synthetic data by translating or paraphrasing texts. Paraphrase the input text to approximately 300 words, aiming to retain the essential information. Here is the text: {INPUT_TEXT}."

B.3. Zero Shot Classification Themes

Prompt text:

"You are tasked with the role of a human annotator, responsible for carefully classifying texts into specific categories related to corporate social responsibility (CSR). Your role involves analyzing the content of various texts, including news articles, reports, and company statements, to identify their alignment with CSR topics. The classification categories are as follows:

1. ENV (Environment): Texts related to environmental sustainability, conservation efforts, impacts of corporate activities on the environment, climate change initiatives, and pollution control.

2. SUP (Sustainable Procurement): Texts discussing sustainable procurement practices, including ethical sourcing, supply chain sustainability, fair trade, and the environmental footprint of products and services.

3. LAB (Labour and Human Rights): Texts detailing labor conditions, human rights issues in business operations, employee welfare, workplace safety, and fair treatment practices within organizations.

4. FBP (Fair Business Practices): Texts focusing on corporate ethics, anti-corruption efforts, transparency, consumer rights, and fair competition in the business landscape.

{INPUT_TEXT}"

Function calling:

```
"functions": [{
    "name": "annotate_text",
    "description": "Analyzes the content of the text, determining its relevance to
        corporate social responsibility topics, and classifies it into one of the
        specified categories",
    "parameters": {
        "type": "object",
        "properties": {
            "text_category": {
                "type": "string",
                "enum": ["ENV", "SUP", "LAB", "FBP"],
                "description": "Corporate social responsibility topic assigned to the text
                    ."
            }
        },
        "required": ["text_category"]
    }
}]
```

B.4. Zero Shot Classification LAB

Prompt text:

"As a human annotator specializing in corporate social responsibility (CSR) with a focus on labor and human rights, your task is to classify texts into detailed categories that reflect various aspects of labor and human rights issues. This role involves a binary relevance classification, meaning for each category listed, you need to decide whether the text is relevant or not. A comprehensive examination of a variety of texts, such as news articles, reports, company statements, and more, is required to identify their relevance to specific topics within the realm of labor and human rights in CSR. A single text may cover multiple aspects of labor and human rights issues, allowing for multiple binary classifications as appropriate:

1. Career Mgmt & Training: Is the text relevant to career development, employee training, and professional growth within organizations?

2. Child Labor, Forced Labor, and Human Trafficking: Does the text address child labor, forced labor, and human trafficking issues?

3. Diversity, Equity, and Inclusion: Is the text focusing on diversity, equity, and inclusion efforts in the workplace?

4. Employee Health & Safety: Does the text concern workplace health and safety policies and practices?

5. External Stakeholder Human Rights: Is the text on human rights issues affecting external stakeholders impacted by corporate activities?

6. Labour Practices and Human Rights: Does the text detail labor practices and human rights considerations within organizations?

7. Social Dialogue: Is the text related to dialogue between employees and management aimed at improving working conditions and relations?

8. Social Discrimination: Does the text deal with social discrimination issues within the workplace or business operations?

9. Working Conditions: Is the text related to employment conditions, such as work hours, pay, and overall work environment?

{INPUT_TEXT}"

Function calling:

```
"functions": [{
    "name": "annotate_text",
    "description": "Analyze text content to determine its binary relevance to labor and
        human rights topics within CSR. For each of the specified categories, the
        annotator will classify the text as either 'relevant' or 'not relevant', based on
        the issues it addresses.",
    "parameters": {
        "type": "object",
        "properties": {
            "text_categories": {
                "type": "object",
                "properties": {
                    "Career Mgmt & Training": {"type": "boolean", "description": "
                        Indicates if the text is relevant to career management and
                        training."},
                    "Child Labor, Forced Labor, and Human Trafficking": {"type": "boolean
                        ", "description": "Indicates if the text addresses child labor,
                        forced labor, and human trafficking issues."},
                    "Diversity, Equity, and Inclusion": {"type": "boolean", "description":
                         "Indicates if the text focuses on diversity, equity, and
                        inclusion efforts."},
                    "Employee Health & Safety": {"type": "boolean", "description": "
                        Indicates if the text is relevant to employee health and safety
                        ."},
                    "External Stakeholder Human Rights": {"type": "boolean", "description
                        ": "Indicates if the text discusses external stakeholder human
                        rights issues."},
                    "Labour Practices and Human Rights": {"type": "boolean", "description
                        ": "Indicates if the text details labor practices and human rights
                         considerations."},
                    "Social Dialogue": {"type": "boolean", "description": "Indicates if
                        the text is related to social dialogue for improving working
                        conditions."},
                    "Social Discrimination": {"type": "boolean", "description": "Indicates
                         if the text deals with social discrimination issues."},
                    "Working Conditions": {"type": "boolean", "description": "Indicates if
                         the text is relevant to working conditions."}
                },
                "required": ["Career Mgmt & Training", "Child Labor, Forced Labor, and
                    Human Trafficking", "Diversity, Equity, and Inclusion", "Employee
                    Health & Safety", "External Stakeholder Human Rights", "Labour
                    Practices and Human Rights", "Social Dialogue", "Social Discrimination
                    ", "Working Conditions"]
            }
        },
        "required": ["text_categories"]
    }
}]
```

C. Contrastive Loss Formulas

Given are vectors α and β, where α contains binary labels indicating whether an in-batch text pair is positive (similar) or negative (dissimilar). Consult Section 4 for a description on positive and negative sample selection. Eq. 1 describes the normalization procedure of the cosine distance values. Eq. 2 denotes the calculation of the contrastive loss, which is the BCE loss between α and β'.

$$\beta' = \text{sig}\left(\frac{\beta}{\Theta}\right) \tag{1}$$

$$\text{CL}(\alpha, \beta) = -\sum i \left[\alpha i \log(\beta' i) + (1 - \alpha i)\log(1 - \beta' i)\right] \tag{2}$$

D. Model Hyperparameters

dataset	tf-idf + SVM	ada-003 + SVM	ada-003 + MLP	ada-003 + LSTM

Themes	ngram = (1,1)
	C = 50
	max iter = 1.000
	max features = 1.000
		C = 10
		max iter = 100
			LR = 1e-4
			n-layers = 2
			n iter = 500
				LR = 1e-3
				epochs = 50
				n-layers = 1
				dropout = 0.1
				hidden dim = 700

ENV	ngram = (1,1)
	C = 10
	max iter = 100
	max features = 10000
		C = 50
		max iter = 1000
			LR = 1e-3
			n-layers = 2
			n iter = 500
				LR = 1e-3
				epochs = 100
				n-layers = 1
				dropout = 0.3
				hidden dim = 700

LAB	ngram = (1,3)
	C = 10
	max iter = 500
	max features = 1000
		C = 100
		max iter = 500
			LR = 1e-3
			n-layers = 2
			n iter = 500
				LR = 1e-3
				epochs = 100
				n-layers = 2
				dropout = 0.3
				hidden dim = 700

Table 3: Optimal hyperparameters for the baseline models, obtained by performing gridsearch experiments.

dataset	DB	DB + CL	DB + DA	DB + CL + DA	RB / BERT	RB / BERT + CL	RB/BERT + DA	RB/BERT + CL + DA	RB-Large	RB-Large + CL	RB-Large + DA	RB-Large + CL + DA
Themes	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 1.0 JI = 1.0 alpha = 0.1	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 1.0 JI = 1.0 alpha = 0.1	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 0.5 JI = 1.0 alpha = 0.2	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 0.5 JI = 1.0 alpha = 0.2	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 0.5 JI = 1.0 alpha = 0.1	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 1.0 JI = 1.0 alpha = 0.2
ENV	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 2e-5 epochs = 10 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 2e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1
LAB	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1

Table 4: Optimal hyperparameters for the LLMs used in this study.

E. Label Names

Idx	ENV	LAB
0	'Air pollution'	'Employee Health & Safety'
1	'Biodiversity'	'Career Mgmt & Training'
2	'Customers Health and Safety'	'Working Conditions'
3	'Energy Consumption & GHGs'	'External Stakeholder Human Rights'
4	'Environmental Services & Advocacy'	'Diversity Equity and Inclusion'
5	'Materials Chemicals & Waste'	'Child Labor Forced Labor and Human Trafficking'
6	'Product End of Life'	'Labour Practices and Human Rights'
7	'Product Use'	'Social Dialogue'
8	'Water'	'Social Discrimination'

Table 5: Label indices and their corresponding names per dataset.

F. Confusion Matrices

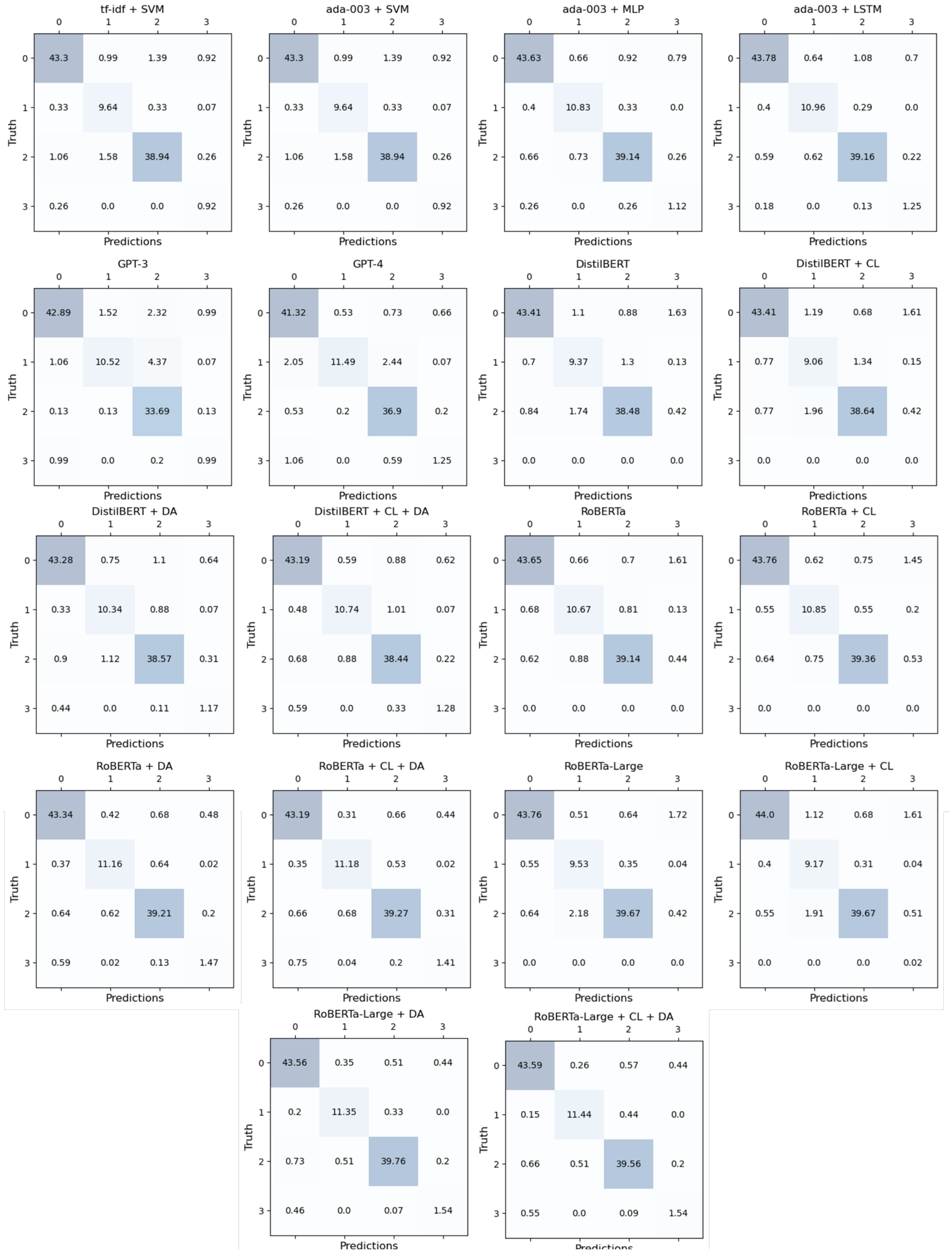

Figure 2: Confusion matrices from all models trained on the themes dataset.

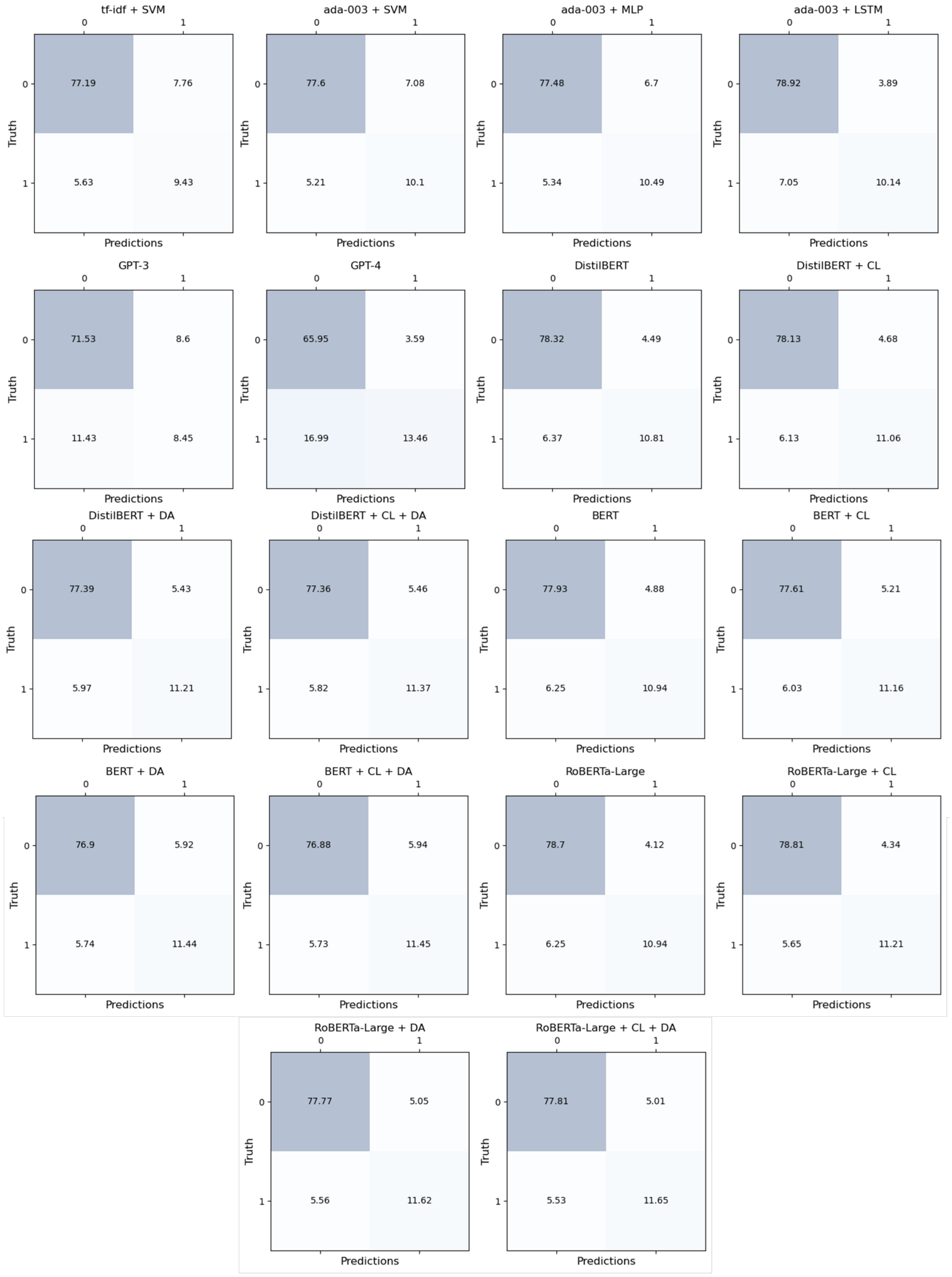

Figure 3: Confusion matrices from all models trained on the ENV dataset.

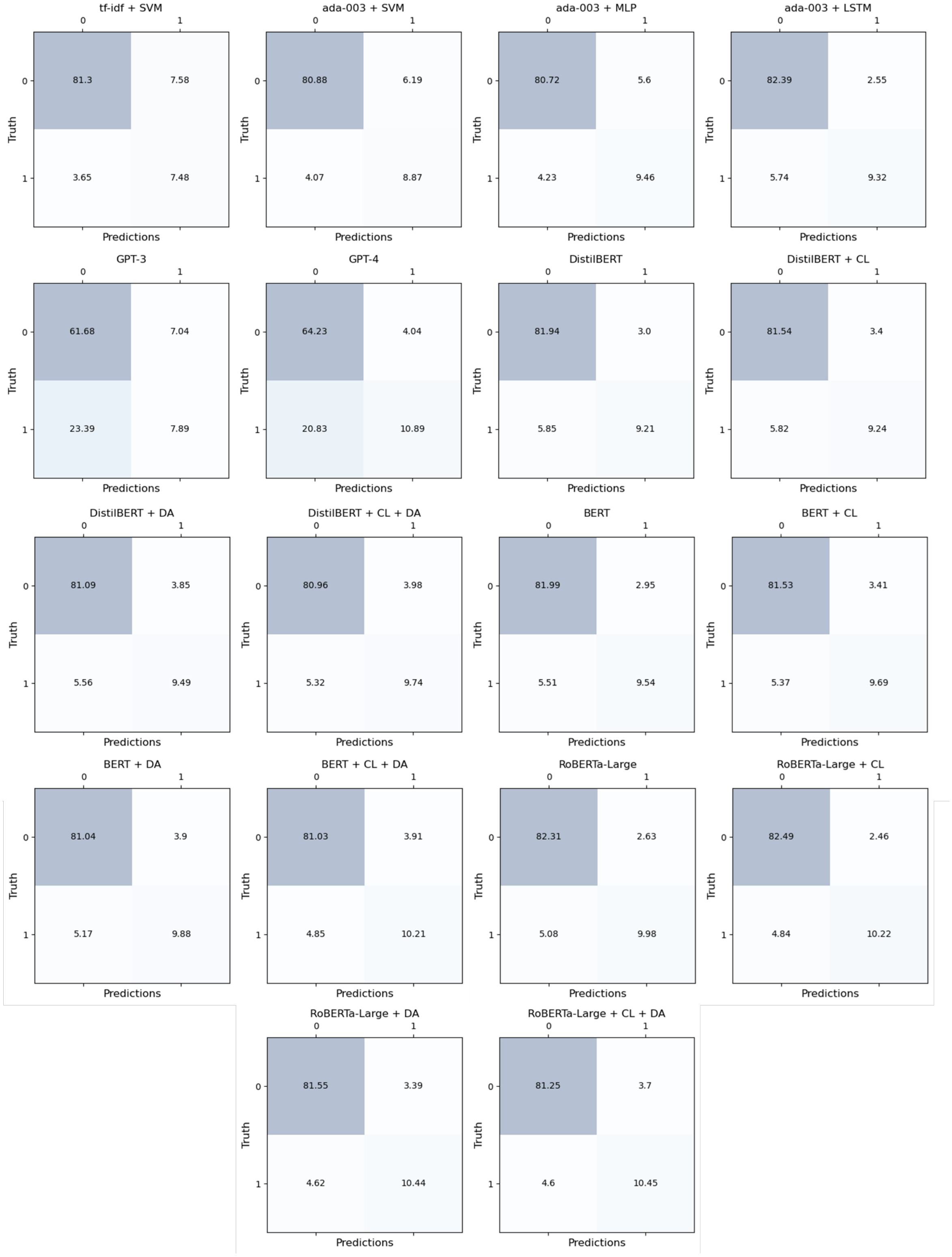

Figure 4: Confusion matrices from all models trained on the LAB dataset.