Improving Cross-Lingual CSR Classification using Pretrained Transformers with Variable Selection Networks and Data Augmentation

Shubham Sharma, Himanshu Janbandhu, Ankush Chopra
Tredence, Whitefield, Bengaluru,India
{shubham.sharma,himanshu.jandbandhu, ankush.chopra}@tredence.com

Abstract

This paper describes our submission to the Cross-Lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics shared task, aiming to identify themes and fine-grained topics present in news articles. Classifying news articles poses several challenges, including limited training data, noisy articles, and longer context length. In this paper, we explore the potential of using pretrained transformer models to classify news articles into CSR themes and fine-grained topics. We propose two different approaches for these tasks. For multi-class classification of CSR themes, we suggest using a pretrained multi-lingual encoder-based model like microsoft/mDeBERTa-v3-base, along with a variable selection network to classify the article into CSR themes. To identify all fine-grained topics in each article, we propose using a pretrained encoder-based model like Longformer, which offers a higher context length. We employ chunking-based inference to avoid information loss in inference and experimented with using different parts and manifestation of original article for training and inference.

Keywords: News Classification, CSR Classification, Multi-lingual Model

1. Introduction

In recent years, Corporate Social Responsibility (CSR) and Sustainability reporting have become significantly important for businesses. Consumers and investors are increasingly conscious of the social and environmental impacts of the companies they engage with or invest in. They often expect businesses to demonstrate a commitment to CSR, sustainability, and ethical business practices. The media plays a crucial role in highlighting CSR issues and holding companies accountable for their actions.

This work is an outcome of a shared task organized in the EcoNLP workshop Nayekoo et al. (2024) which had two subtasks. These subtasks focus on automatically identifying themes and topics present in news articles. The first subtask involves multi-class classification of multi-lingual news articles. We experimented with various multi-lingual encoder-only transformer models. We also experimented with various transformations of input text such as translation, summarization and chunking for accurately classifying the news articles. The best performing model was mDeBERTa (He et al., 2023) in conjunction with variable selection network trained on article text and the titles.

The second subtask is a multi-label classification for fine-grained labels among two CSR themes: ENV and LAB. This subtask was conducted with only English language news articles. The best performing model for this task was Longformer (Beltagy et al., 2020) with multi-label head and custom threshold for each class head.

2. Related Work

The field of multi-lingual text classification, integral to natural language processing (NLP), has seen significant advancements with the introduction of pre-trained transformer-based models such as mBERT.Gürel and Emin (2021) have effectively utilized mBERT for multi-lingual text classification, showcasing its proficiency across various languages. Complementing this, Pujari et al. (2021) have explored the use of a transformer-based multi-task model for multi-label classification, employing a novel approach that trains a dedicated classifier for each node by merging transformers with hierarchical algorithms.

In a related development, Wang et al. (2021) have adopted graph convolutional networks (GCN) for cross-lingual text classification. Their method involves constructing a heterogeneous graph where documents and words serve as nodes. These nodes are interconnected through a network of relationships defined by part-of-speech roles, semantic similarity, and document translations, facilitating a comprehensive understanding of language nuances.

Parallel to these developments, the area of Corporate Social Responsibility (CSR) topic identification has been flourishing. Chae and Park (2018) have applied a probabilistic topic modeling-based computational text analysis framework to examine the prevalence, evolution, and correlation of CSR topics. Further advancing the research, Salvatore et al. (2022) have employed a structural topic model to detect CSR topics and assess the impact of time

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 306–318

and sector on the proportion of discussions surrounding these topics.

3. Dataset and Problem statements

3.1. Data Scraping

The datasets provided for both subtasks consisted of URLs of the articles, using which we had to scrape text. To accomplish this, we utilized a Python library called Newspaper3K (Ou-Yang, 2018), which provided us with the text of the articles (excluding boilerplate), article titles, and other metadata. This information was extracted from the HTML of the URLs by the library itself. Due to challenges in web scraping, we were unable to scrape all instances provided in the original data of both the tasks.

3.2. Subtask A

This subtask aims to identify the theme of news articles, which could be in English, French, or simplified Chinese. The themes are categorized into four labels: environment, labor and human rights, fair business practices, and sustainable procurement. Therefore, it becomes a multi-lingual and multi-class classification problem.

3.2.1. Dataset

In our experiments, we utilized the Corporate Social Responsibility theme recognition dataset provided by the event organizers. This dataset included URLs and CSR themes for each news article. Since the dataset did not contain the actual text of the news articles, we resorted to web scraping to extract it. The four main themes were categorized as follows: 1. ENV (Environment), 2. LAB (Labor and Human Rights), 3. FBP (Fair Business Practices), and 4. SUP (Sustainable Procurement).
The provided dataset was only available in English, but predictions needed to be made in three languages: English, French, and simplified Chinese. To deal with the small sample size, we also employed a synthetic data generation method for the SUP class as described in section 4.4. The class distribution of the data is provided in Table 1. The tokenized length distriution of scraped articles is provided in Table 2.

Data	ENV	LAB	FBP	SUP
Original Data	706	652	197	39
Data Post Scraping	633	547	179	33
With Synthetic Data	633	547	179	153

Table 1: Label Distribution

Max	Min	Mean	Std
13876	23	566	712

Table 2: Length Distribution in scraped articles for Task A.

3.3. Subtask B

This subtask aims to classify fine-grained CSR topics within the Environment (ENV) and Labor and Human Rights (LAB) themes (English), allowing for multiple topics to be assigned to an article within each specified theme. Therefore, it presents a multi-label multi-class classification challenge.

3.3.1. Dataset

For this subtask, we received two separate datasets for ENV and LAB, each with its own set of labels. These datasets were also scraped as discussed in section 3.1. The label distribution before and after scraping for the ENV dataset is provided in Table 4. The tokenized length distribution of scraped articles is provided in Table 3.

Dataset	Max	Min	Average	Std
ENV	4866	35	539	542
LAB	11867	36	655	1256

Table 3: Length Distribution in scraped articles for Task B.

Similarly, for the LAB class, the label distribution before and after scraping is detailed in Table 5. We observed that the labels "External Stakeholder Human Rights" and "Social Discrimination" had a smaller number of training examples provided. Additionally, they were not highly co-occurring with other classes. Consequently, we created artificial data for these two labels using the method described in section 4.4.

ENV	Original Data	Records Post Scraping
Air Pollution	36	31
Biodiversity	62	51
Customer Health and Safety	62	48
Energy Consumption and GHGs	366	313
Environmental Services and Advocacy	242	204
Materials, Chemical and Waste	112	92
Product End of Life	73	64
Product Use	44	36
Water	71	57

Table 4: Number of records for ENV dataset

LAB	Original Data	Records Post Scraping	With Synthetic Data
Career Mgmt and Training	77	42	42
Child Labor, Forced Labor, and Human Trafficking	7	4	4
Diversity, Equity, and Inclusion	149	99	99
Employee Health and Safety	138	100	100
External Stakeholder Human Rights	14	12	36
Labour Practices and Human Rights	47	29	29
Social Dialogue	52	28	28
Social Discrimination	18	18	42
Working Conditions	201	128	128

Table 5: Number of records for LAB dataset

4. Methodology

Both subtasks involve classification, making encode-only transformer models a suitable choice for fine-tuning. We opted for multi-lingual models for subtask A and English language models for subtask B. Experimentation with various models, some of which had a maximum sequence length of 512, was conducted. However, a significant portion of the input news articles exceeded the 512-length limit, as indicated in Table 2 and Table 3. Thus, we explored methods for handling such documents, including the use of models like Longformer with higher sequence lengths and the employment of data transformation through chunking and summarization.

We used full finetuning of the transformer-based model with Pytorch library with appropriate cross-entropy losses for both the subtasks. We used three models, DeBERTa, MDeBERTa and Longformer for fine-tuning.

Data augmentation, transformation, and experimentation with model architecture were key components in our search for optimal solutions. We designed numerous experiments by exploring various combinations of these steps. The specifics of the data and model architecture-related variations are outlined below

4.1. Summarization

Abstractive summarization was utilized to generate summaries for the articles. This approach enabled us to reduce the length of the articles to fit within the BERT model's context length limit. The articles were segmented into paragraphs, ensuring they did not exceed the maximum context size of the summarization model. Each paragraph was then summarized to a specified number of words, depending on the number of chunks, to ensure the summarized article fit within the context length limit of the downstream classification model. We utilized a T5-based (Raffel et al., 2023) summarization model called "Falconsai/text_summarization", (See Appendix 8.5) which is hosted on Hugging Face (Wolf et al., 2020).

4.2. Chunking

To address the challenge posed by excessively long articles in our dataset, we attempted to divide the dataset into multiple sequences or chunks. Each chunk was created not to surpass the maximum context length limit of the model utilized during the experiment. In subtask A, we could allocate each chunk to the same category and utilize this data to train the model.

4.2.1. Mean Probability based Prediction.

This approach involved leveraging probabilities from each chunk to calculate the overall average probability for each class. Subsequently, we utilized this mean probability to make predictions.

4.2.2. Max Voting based Prediction

This method was solely employed for subtask A. Each chunk predicted a class based on its predicted probabilities. These chunk-wise predictions were aggregated as votes for each class, and the predicted class was determined by the maximum number of chunks voting for it. In case of a tie, we used the maximum probability of each class to make the final prediction.

4.3. Variable Selection Network (VSN)

Given the longer length of the texts, we conjectured that the CLS token might not always encapsulate the complete meaning and context of the input. This prompted us to explore methods to provide selected additional information to the classification layer for improved predictions. VSN (Lim et al., 2020) is a neural network architecture designed for feature selection in high-dimensional datasets. It utilizes a gating mechanism to learn the most relevant features for a given task, focusing on important features while disregarding irrelevant ones. The architecture of VSN is illustrated in Figure 1.

We employed VSN to generate an embedding representing all token embeddings in the BERT-based model. This embedding was concatenated with the CLS token and passed for fine-tuning. (We refer to Figure 6 which illustrates the VSN architecture.)

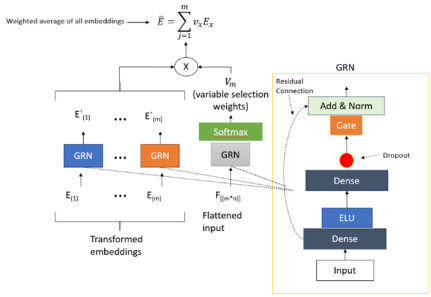

Figure 1: Variable Selection Network

4.4. Synthetic Data Generation

We utilized the GPT-4 (OpenAI et al., 2024) model by OpenAI to generate artificial samples of minority classes. This was achieved using prompts specifically tailored for generating synthetic text articles. The prompts utilized for generating data for selected classes in both subtasks can be found in appendix 8.4.

4.5. Title Concatenation

As the news articles provided by the organizers were in URL format, we had access to rich metadata from these news webpages. After careful analysis and observation of available common metadata fields, we opted to incorporate page titles along with article text. Page titles are crafted to succinctly represent page content and offer high-quality condensed information.
In certain experiments, we concatenated the title with the original text of the news article, separated by the SEP token of the model. Different token type IDs were assigned to both the title and text to ensure that the model could attend to both sequences.

4.6. Translation

The data provided by the organizers for subtask A was in English, while the test set was announced to be in French and Chinese, in addition to English. Given the challenge of identifying, scraping, and tagging relevant French and Chinese data, we decided to translate the provided English articles into French and Chinese for training. We ensured consistency by maintaining the same records in the validation set across all languages. We utilized pretrained models from Hugging Face for translation, specifically "Helsinki-NLP/opus-mt-tc-big-en-fr" (See Appendix 8.5) for French translation and "Helsinki-NLP/opus-mt-en-zh" (See Appendix 8.5) for simplified Chinese translation. (Tiedemann and Santhosh, 2020).

4.7. Dynamic Weighted Loss

As the dataset was imbalanced, we decided to assign weights to each class in the loss function. We started with equal weights for each class in the initial epoch, then at the start of the next epoch, we assigned weights in such a way that class with minimum f1 should get maximum weight. Here is the formula:
Weight for class i,

$$W_i = \frac{1-F1_i}{\sum_{j=0}^{n}(1-F1_j)}$$

Where n is the number of classes

5. Experiments

After implementing the methods described in the previous section, we conducted numerous experiments. In this section, we present our best systems for both subtask A and subtask B. For details on other experiments, please refer to the Appendix. (refer 8.1, 8.2, 8.3)

5.1. Subtask A

As outlined in section 3.2, subtask A poses a multilingual multi-class classification challenge. Through a combination of methodologies discussed in section 4, we discovered that utilizing article titles indeed improves performance. In addition, we observed that using variable selection network along with title concatenation further enhanced the performance of the system.

ENV F1	LAB F1	FBP F1	SUP F1	ACC	language
96.3	96.5	95.4	71.4	94.9	English
96.8	96.6	94.5	79.1	95.3	French
92.4	92.6	86.6	66.6	90.0	Chinese

Table 6: Subtask A validation set results per language

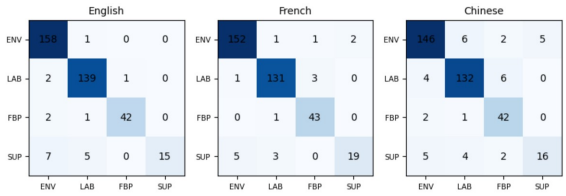

Figure 2: : Subtask A confusion matrix per language

Air pollution	Biodiversity	Customers Health and Safety	Energy Consumption and GHGs	Environmental Services and Advocacy	Materials, Chemicals and Waste	Product End of Life	Product Use	Water	micro avg
71.42	66.66	85.71	89.69	60.93	59.45	68.75	37.50	72	73.20

Table 7: Subtask B, ENV, F1-Scores

Career Mgmt and Training	Child Labor, Forced Labor, and Human Trafficking	Diversity, Equity, and Inclusion	Employee Health and Safety	External Stakeholder Human Rights	Labour Practices and Human Rights	Social Dialogue	Social Discrimination	Working Conditions	micro$_{avg}$
69.56	66.66	72.22	81.63	40	22.22	94.11	57.14	80.48	71.48

Table 8: Subtask B, LAB, F1-Scores

We evaluated our models based on their F1 scores for each class and language. Our best-performing model utilized the DeBERTa-based multi-lingual model, specifically the "microsoft/mdeberta-v3-base (8.5)" model hosted on Hugging Face and pretrained on CC100 multilingual data. This best system incorporated artificial data for the SUP class, as discussed in section 4.4, and utilized translated datasets, as described in section 4.6. Additionally, it involved title concatenation, as outlined in section 4.5, and employed a variable selection network, as discussed in section 4.3. In terms of inference, we utilized chunking with mean probability, as detailed in section 4.2.1 . During model training, we included original SUP data in the validation set and artificial SUP data in the training set. For classes other than SUP, original data were used in training and validation set. The results of the best-performing model on validation set are provided in Table 6. In Figure 2, i-th row and j-th column entry indicates the number of samples with true label being i-th class and predicted label being j-th class. The results of best performing model on test set are given in the Table 12.

5.2. Subtask B

In subtask B, a multi-class, multi-label problem was tackled using fine-tuning of transformer-based models, employing the binary-crossentropy loss for each class. This approach suits multi-label classification due to its capability to handle instances with multiple labels and provide a probabilistic interpretation of predictions, facilitating overall loss reduction.
In the ENV dataset, shorter articles were predominant compared to the LAB dataset (see Table 3). Initially using DeBERTa, the baseline model was shifted to Longformer due to a large portion of data exceeding 512 tokens, the maximum context size for DeBERTa. Title concatenation notably enhanced performance across experiments with Longformer. Further experiments included VSN exploration, adjusting context-length limits with Longformer, and prediction with chunked articles, though no improvement over the best model was observed. The optimal model for the ENV dataset was Longformer with a 1500 context-length limit and title concatenation.
Similarly, in the LAB dataset, Longformer outperformed DeBERTa after summarization. Title concatenation, however, led to performance deterioration. The same experiments as in the ENV dataset were conducted, with the best model again being Longformer with a 1500 context-length limit. Best results for ENV and LAB subtask on validation set can be found Table 7 and Table 8 in respectively. Summarized results of Task B on test set can be found in Table 13.

6. Conclusion

In this paper, we introduce systems developed by our team for the Cross-Lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics tasks. This task aims at automatically classifying news articles into CSR themes and identifying fine-grained topics within them. To achieve this, we finetuned transformer models along with a variable selection network to classify articles into suitable CSR themes. By experimenting with title along with article text we uncovered that using metadata effectively along with the article text can help immensely in improving the accuracy of the classification.

7. Bibliographical References

Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Bongsug (Kevin) Chae and Eunhye (Olivia) Park. 2018. Corporate social responsibility (csr): A survey of topics and trends using twitter data and topic modeling. *Sustainability*, 10(7).

Alaeddin Gürel and Emre Emin. 2021. ALEM at CASE 2021 task 1: Multilingual text classification on news articles. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 147–151, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. 2020. Temporal fusion transformers for interpretable multi-horizon time series forecasting.

Yola Nayekoo, Sophia Katrenko, Véronique Hoste, Aaron Maladry, and Els Lefever. 2024. Shared task for cross-lingual classification of corporate social responsibility (csr) themes and topics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Niko-

las Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Lucas Ou-Yang. 2018. Newspaper3k: Article scraping curation. Technical report.

Subhash Chandra Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. A multi-task approach to neural multi-label hierarchical patent classification using transformers. In *Advances in Information Retrieval*, pages 513–528, Cham. Springer International Publishing.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Camilla Salvatore, Silvia Biffignandi, and Annamaria Bianchi. 2022. Corporate social responsibility activities through twitter: From topic model analysis to indexes measuring communication characteristics. *Social Indicators Research*.

Jörg Tiedemann and Thottingal Santhosh. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. Cross-lingual text classification with heterogeneous graph neural network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

8. Appendix

8.1. Table A: Experiments and results on Subtask A

Experiment No.	1	2	3	4	5	6
MODEL_NAME	mDeBERTa	mBERT	Longformer	mDeBERTa	mDeBERTa	mDeBERTa
Methods_used				4.2, 4.2.2	4.2, 4.2.1	4.1
English						
ENV F1	97.79	94.67	92.77	94.26	95.95	93.04
LAB F1	96.48	94.04	93.19	94.55	95.74	94.48
FBP F1	91.67	89.16	93.18	86.67	92.47	82.35
SUP F1	0.00	35.29	0.00	60.00	60.00	40.00
ACCURACY	95.45	92.33	92.07	92.92	94.90	91.22
French						
ENV F1	91.67	89.76	90.21	87.36	90.96	89.47
LAB F1	92.26	85.42	90.04	87.45	88.30	89.11
FBP F1	86.36	61.11	81.32	72.94	80.00	77.11
SUP F1	0.00	16.67	0.00	20.00	22.22	37.50
ACCURACY	90.34	83.81	88.10	84.70	87.54	86.69
Chinese						
ENV F1	91.86	91.19	90.96	93.25	92.26	93.46
LAB F1	90.91	90.58	86.86	88.97	91.37	90.68
FBP F1	86.02	77.65	77.42	74.73	81.63	82.93
SUP F1	0.00	28.57	0.00	0.00	0.00	53.33
ACCURACY	89.77	88.07	86.69	87.54	89.52	90.27

Experiment No.	7	8	9	10	11	12	**13**
MODEL_NAME	mDeBERTa	mDeBERTa	mDeBERTa	mDeBERTa	mDeBERTa	mDeBERTa	**mDeBERTa**
Methods_used	4.4	4.2.1, 4.4	4.7	4.2.1, 4.4, 4.6	4.2.1, 4.4, 4.5, 4.6	4.2.1, 4.3, 4.4, 4.6	**4.2.1, 4.3, 4.4, 4.5, 4.6**
English							
ENV F1	93.05	93.05	94.05	94.41	95.71	93.46	**96.34**
LAB F1	95.47	95.50	95.31	95.95	95.92	95.04	**96.53**
FBP F1	94.12	87.50	89.16	93.02	89.16	90.53	**95.45**
SUP F1	65.12	69.57	0.00	70.37	69.77	70.00	**71.43**
ACCURACY	92.40	91.90	92.80	93.13	93.50	91.80	**94.90**
French							
ENV F1	89.51	93.01	89.22	92.95	95.60	92.60	**96.82**
LAB F1	86.93	93.53	88.65	95.59	94.70	94.38	**96.68**
FBP F1	73.97	86.08	79.49	90.32	85.71	87.23	**94.51**
SUP F1	60.47	57.89	0.00	68.09	78.26	73.08	**79.17**
ACCURACY	85.25	90.60	86.60	91.90	93.09	91.10	**95.30**
Chinese							
ENV F1	89.41	88.24	91.99	88.52	92.11	91.03	**92.41**
LAB F1	91.16	92.65	91.64	93.43	94.24	92.78	**92.63**
FBP F1	85.00	87.18	86.75	84.21	88.37	84.21	**86.60**
SUP F1	31.25	25.00	22.22	50.00	66.67	69.23	**66.67**
ACC	87.13	86.90	90.30	87.20	90.80	89.19	**90.00**

Table 9: All experiments results for Subtask A

8.2. Table B : Subtask B, Experiments on ENV dataset

Experiment	1	2	3	4	**5**	6	7
Model-Name	DeBERTa	Longformer	DeBERTa	DeBERTa	**Longformer**	DeBERTa	DeBERTa
Context Length	512	1500	512	512	**1500**	512	512
Methods Used			4.1	4.5	**4.5**	4.3	4.3, 4.5
Air pollution f1-score	50.00	50.00	50.00	33.33	**71.43**	52.63	50.00
Biodiversity f1-score	12.50	50.00	40.00	33.33	**66.67**	55.56	62.07
Customers Health and Safety f1-score	58.33	66.67	58.33	63.64	**85.71**	76.92	80.00
Energy Consumption & GHGs f1-score	86.08	86.08	84.81	87.65	**89.70**	87.18	87.43
Environmental Services & Advocacy f1-score	42.55	44.68	46.81	56.86	**60.94**	58.06	59.63
Materials, Chemicals & Waste f1-score	48.15	51.85	44.44	40.00	**59.46**	56.76	53.57
Product End of Life f1-score	56.25	43.75	64.52	62.50	**68.75**	68.57	68.57
Product Use f1-score	0.00	42.86	14.29	40.00	**37.50**	16.67	34.04
Water f1-score	64.52	81.25	56.25	64.00	**72.00**	76.47	66.67
micro_avg	58.98	64.41	61.54	64.72	**73.20**	68.05	66.20

Experiment	8	9	10	11	12
Model Name	Longformer	Longformer	Longformer	DeBERTa	Longformer
Context Length	1500	1500	2000	512	1200
Methods used	4.3	4.3, 4.5	4.5	4.2.1, 4.5	4.5
Air pollution f1-score	66.67	70.00	57.14	61.54	71.43
Biodiversity f1-score	77.78	71.43	66.67	25.00	63.64
Customers Health and Safety f1-score	83.33	84.21	80.00	77.78	73.68
Energy Consumption & GHGs f1-score	88.89	85.53	91.12	85.71	89.17
Environmental Services & Advocacy f1-score	57.69	60.00	59.35	60.00	59.35
Materials, Chemicals & Waste f1-score	54.29	51.16	57.63	43.24	50.00
Product End of Life f1-score	71.43	74.29	65.00	70.59	70.27
Product Use f1-score	29.41	24.00	27.59	35.29	28.57
Water f1-score	88.24	66.67	81.82	55.17	80.00
micro_avg	69.17	68.56	69.96	64.92	67.30

Table 10: All experiments results, Subtask B, ENV Dataset

8.3. Table B : Subtask B, Experiments on LAB dataset

Experiment	**1**	2	3	4	5	6
Model Name	**Longformer**	DeBERTa	Longformer	DeBERTa	DeBERTa	DeBERTa
Context Length	**1500**	512	1500	512	512	512
Methods Used		4.5	4.5	4.1	4.3	4.3, 4.5
Career Mgmt & Training f1-score	**69.56**	60.86	60.86	60.00	76.19	63.15
Child Labor, Forced Labor, and Human Trafficking f1-score	**66.66**	0	13.33	3.33	14.28	2.77
Diversity, Equity, and Inclusion f1-score	**72.22**	66.67	66.67	68.75	66.67	76.47
Employee Health & Safety f1-score	**81.63**	81.82	72.34	69.57	72.34	72.73
External Stakeholder Human Rights f1-score	**40.00**	14.29	50.00	6.82	12.24	23.08
Labour Practices and Human Rights f1-score	**22.22**	37.50	32.43	20.00	23.53	34.78
Social Dialogue f1-score	**94.12**	80.00	80.00	80.00	80.00	72.73
Social Discrimination f1-score	**57.14**	38.10	25.00	54.55	50.00	57.14
Working Conditions f1-score	**80.49**	78.95	85.37	81.72	81.93	77.50
micro_avg	**71.49**	48.56	63.94	44.50	58.98	51.38

Experiment	7	8	9	10
Model Name	Longformer	Longformer	Longformer	Longformer
Context Length	1500	1500	1500	1200
Methods Used	4.3	4.3, 4.5	4.4	
Career Mgmt & Training f1-score	81.81	63.15	66.66	70
Child Labor, Forced Labor, and Human Trafficking f1-score	100	12.50	100	66.66
Diversity, Equity, and Inclusion f1-score	68.42	70.00	70.59	71.79
Employee Health & Safety f1-score	76.92	82.61	80.85	72.34
External Stakeholder Human Rights f1 score	66.67	25.00	40.00	24.00
Labour Practices and Human Rights f1-score	17.54	37.04	17.78	24.24
Social Dialogue f1-score	87.50	88.89	87.50	94.12
Social Discrimination f1-score	75.00	42.11	57.14	53.33
Working Conditions f1-score	78.16	86.42	79.55	81.08
microavg	65.28	64.77	66.42	64.47

Table 11: All experiments results, Subtask B, LAB Dataset

8.4. PROMPTS

As mentioned in the section 4.4 we generated synthetic data using gpt4 . The prompt that we used for generating those synthetic samples is given below:

Sample Artificial Data related to Sustainable Procurement:
Greenway Partners Join Forces with Financing Institutions to Promote Sustainable Production In a bold move towards fostering a greener future, tech giant Greenway Partners has teamed up with several financing entities to reward their suppliers who work towards inclusivity and sustainability. Suppliers would have to prove their commitment towards environmentally friendly practices to get special financing rates. The program's evaluation basis includes an independently made roadmap and classification outline designed in tandem with GOSE, an international charity organization focused on global environmental disclosure. The initiative is expected to incentivize suppliers to cut down their carbon emission, helping Greenway Partners meet its emissions targets. Greenway's collaboration is part of its broader effort to assist clients in achieving their own eco-friendly aspirations. We're thrilled to be a part of Greenway's vision in realizing a sustainable future for all. We firmly believe in achieving net zero emission and we're more than happy to assist Greenway in their significant emissions reduction's strategies says Katy Peterson, director of sustainability programs at one of the financing institutions. Senior corporate banking executive Josh Crawford adds, 'This venture further solidifies our long-term association with Greenway that spans over years and across multiple countries.'

Instruction: You are given an article related to sustainable procurement. You have to generate 2 artificial articles related to sustainable procurement similar to given article. Make sure article should not contain information from below given article.

Article:

{article}

format instructions: {format_instructions}

Figure 3: : Prompt for Sustainable Procurement data

Sample Artificial Data related to External Stakeholder human rights:
In the context of a globally intertwined business environment, the rights and interests of players that extend beyond immediate business circles hold undeniable relevance. These rights often encapsulate diverse aspects that organically pertain to the ethos of international social justice and ethical operations. For example, a company operating across international territories must ensure that the safety and welfare of its workforce is upheld irrespective of its operational decisions. This signifies the company's commitment to creating an inclusive work culture that is driven by diversity and empowerment. On similar lines, corporations have a responsibility towards limiting any adverse environmental ramifications stemming from business operations. This commitment, though presents its challenges, resonates with a higher ethos of global health and sustainability. These concepts shaping the dynamics between businesses and international societal players pave the way towards an equitable environment that drives companies towards innovative and sustainable practices yielding mutual benefits. Therefore, it is the shared responsibility of companies to acknowledge and prioritize these diverse interests of players outside immediate business relationships for the growth of inclusive societies and robust businesses.

Instruction: You are given an article related to external stakeholder human rights. You have to generate 2 artificial articles related to external stakeholder human rights. Make sure article should not contain information from below given article and avoid mentioning directly human rights or external stakeholder.

Article:

{article}

format instructions: {format_instructions}

Figure 4: Prompt for External Stakeholder Human Rights

Sample Artificial Data related to Social Discrimination:

Employment violations related to unfair expectations have surfaced in a recent case involving EnviroLibrium, a leading environmental solutions company headquartered in Houston, Texas. The company is facing accusations from multiple employees who claim it employs unfair practices when addressing employees with certain health conditions. These employees leveled allegations stating that their employer imposed disconnected, counterproductive, and invasive requirements on them, including forced meetings with medical professionals, regardless of their consent or personal medical treatment plans. These behaviors directly oppose fair labor principles and may infringe on employee rights to receive reasonable accommodation for their medical conditions. This case represents the need for employers to comprehend the complexities surrounding employee health conditions and the provision of adaptive support in compliance with their rights. EnviroLibrium has agreed to provide comprehensive training on Americans with Disabilities Act (ADA) to educate its workforce, to avoid such potential missteps in the future. The company, however, denies any allegations of discriminatory practices, asserting their commitment to a diverse and inclusive work environment. This case further emphasizes the necessity of abiding by ADA guidelines, not just towards safeguarding employee rights, but also towards cultivating an inclusive work culture.

Instruction: You are given an article related to social discrimination. You have to generate 2 artificial articles related to social discrimination. Make sure article should not contain information from below given article and avoid mentioning directly social discrimination.

Article:

{article}

format instructions: {format_instructions}

Figure 5: Prompt for Social Discrimination

8.5. Huggingface Model References

'microsoft/deberta-v3-base' :https://huggingface.co/microsoft/deberta-v3-base
'microsoft/mdeberta-v3-base' :https://huggingface.co/microsoft/mdeberta-v3-base
'allenai/longformer-base-4096' :https://huggingface.co/allenai/longformer-base-4096
'Falconsai/text_summarization' :https://huggingface.co/Falconsai/text_summarization
'Helsinki-NLP/opus-mt-en-fr' :https://huggingface.co/Helsinki-NLP/opus-mt-en-fr
'Helsinki-NLP/opus-mt-en-zh' :https://huggingface.co/Helsinki-NLP/opus-mt-en-zh

8.6. Test Set Results

acc.	prec.	rec.	f1-m	f1 -w	language
0.95	0.97	0.95	0.77	0.96	English
0.78	0.87	0.78	0.61	0.81	Chinese
0.94	0.95	0.94	0.87	0.94	French

Table 12: Subtask A test set results per language

task	accuracy	precision	recall	f1 macro	f1 weighted	hamming
ENV	87.53	88.40	87.53	72.48	87.67	87.53
LAB	87.54	90.12	87.54	68.24	87.93	87.54
overall	87.54	89.26	87.54	70.36	87.80	87.54

Table 13: Subtask B, test set results

8.7. Illustrations

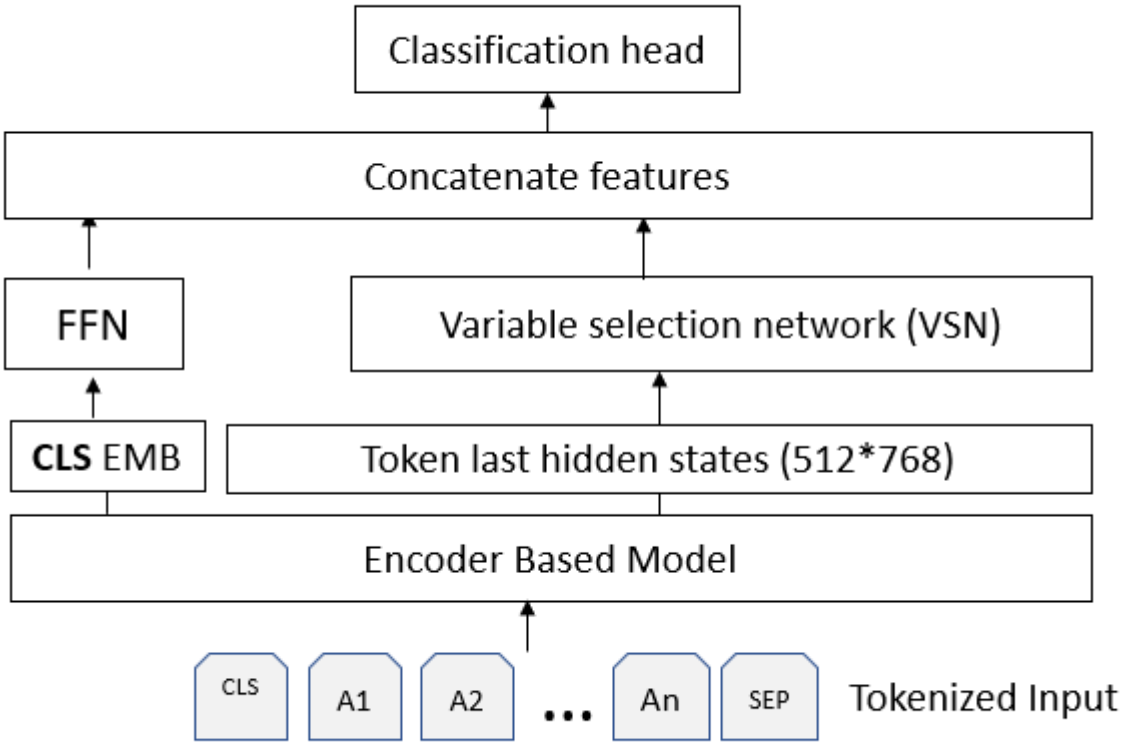

Figure 6: : CLS and VSN Architecture