Jetsons at the FinNLP-2023: Using Synthetic Data and Transfer Learning for Multilingual ESG Issue Classification

Parker Glenn*, **Alolika Gon***, **Nikhil Kohli***, **Sihan Zha***, **Parag Pravin Dakle**, **Preethi Raghavan**
Fidelity Investments, AI Center of Excellence
{parker.glenn, alolika.gon, nikhil.kohli, sihan.zha, paragpravin.dakle, preethi.raghavan}@fmr.com

Abstract

In this paper, we describe the various approaches by the *Jetsons* team for the Multilingual ESG Issue Identification Task (ML-ESG) to classify articles into ESG (environmental, social, and corporate governance) issues they are related to. For English and French articles, we finetune multilingual BERT with synthetic data in a single-label classification setting. For the Chinese articles, we employ transfer learning to leverage the full breadth of the multilingual training data. Our methods achieve 1st place on the leaderboard for French, and 5th place for both English and Chinese.

1 Introduction

ESG (environmental, social, and governance) investing introduces a set of standards to judge investments by values corresponding to specific issues. Examples of these issues include "Chemical Safety", "Controversial Sourcing", and "Carbon Emissions". The International Joint Conferences on Artificial Intelligence (IJCAI) shared task (Chen et al., 2023) presents a fine-grained multilingual classification task based on a taxonomy of these ESG issues.

We approach this task using several strategies, including 1) transfer learning to the multi-label Chinese data using mBERT (Devlin et al., 2018), 2) augmentation with synthetic data generated with LLMs in zero-shot and few-shot settings, and 3) T5 variants (Xue et al., 2021, 2022) for multiclass text classification.

2 Related work

Language models have been used for various financial tasks like named-entity recognition, sentiment analysis, or document classification. Previous works have performed domain-specific pre-training of language models for different financial tasks (Araci, 2019; Huang et al., 2022; Shah et al., 2022; Lu et al., 2023; Wu et al., 2023). However, until recently, only a few works have explored using language models for ESG-related tasks. Raman et al. 2020 evaluate the impact of using embeddings generated by language models on the classification of sentences concerning their relevance to the ESG domain. Mehra et al. 2022 pre-train a BERT model on ESG-related text to show improvement on classification tasks. Nugent et al. fine-tune an English BERT-style model on an ESG document classification dataset and evaluate using data generation as an augmentation strategy.

Kær Jørgensen et al. 2021 extend the idea of pre-training on financial text to multilingual text and evaluate different sentence classification tasks in seven languages. Jørgensen et al. 2023 evaluate various language models on a multilingual financial topic classification dataset to highlight areas of improvement for low-resource languages.

The work of Nugent et al. 2021 is closest to work presented in the paper. The authors use the back-translation task to generate additional input data. This work, however, performs ESG document classification in a mono- and multilingual setting. Additionally, we use a large language model to generate additional data using just the ESG topic compared to performing back-translation.

3 Data

We use the dataset described in Chen et al. (2023) for this task. The training dataset consists of articles in three languages: English (en), French (fr), and Chinese (zh). Alongside the articles are the corresponding ESG issues these articles are related to. The English and French dataset each contains about 1200 articles. These datasets are single-labeled with one out of 35 ESG classes, as designated by the MSCI[1]. The Chinese training set

*These authors contributed equally to this work

[1]https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map

Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP) and 2nd Multimodal AI For Financial Forecasting (Muffin), Macao, August 20, 2023.

contains 996 articles. The Chinese labels merge the MSCI classes with those designated by the SASB[2] for a total of 46 total labels. These Chinese data points are multilabel, and each article is classified with a minimum of one and a maximum of 13 labels. Figure 1 shows the distribution of training and validation instances on the 24 most popular classes.

3.1 Synthetic Data Generation with large language models

We leverage the power of open-source large language models (gpt-3.5-turbo[3]) in generating text for augmenting the dataset to improve the class imbalance. For all three languages (English, French, and Chinese), given an ESG label, we generate 'News Title' and 'News Summary' for each instance. We generate a total of 413 data points for 11 different labels. We choose these labels based on the class-wise performance metrics and class distribution.

We categorize the ESG labels into two categories - ambiguous and non-ambiguous. Here we define ambiguity as a label being open to more than one interpretation and requiring some domain expertise to resolve the ambiguity. We employ two different strategies for generating samples for these two categories. For non-ambiguous topics, we use zero-shot generation. For ambiguous topics, we use few-shot generation to ensure that the generated samples are related to the ESG domain. Below is an example of a zero-shot prompt.

```
Give 10 examples of news related
to ESG (Environmental, Social, Governance)
topic 'Electronic Waste'. Each example
should have a news title, news summary and
tags related to the article. Generate
these examples in french language.
```

4 Models

4.1 Chinese

Given the disjoint task setup of the Chinese data with the English and French data (multilabel vs. multiclass, respectively), it is difficult to train a single multilingual clasifier for all three languages. To utilize the value of English and French data, we adopt a transfer learning technique to train a model

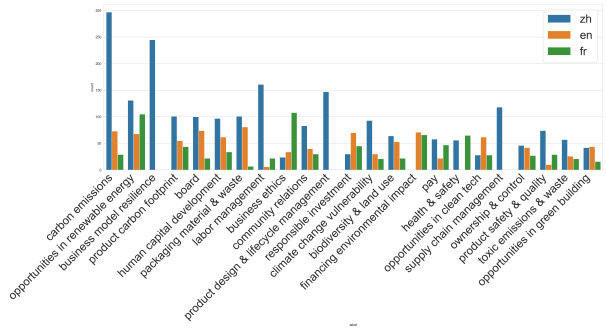

Figure 1: Frequency of the most popular 24 labels in the Train and Validation splits across the 3 languages.

to classify Chinese texts. We use the bert-base-multilingual-cased[4], which is based on the BERT model (Devlin et al., 2018), as our base model to reconcile different languages. We first fine-tune the base model using the English and French data to do single-class classification. Subsequently, we further fine-tune the model from the previous steps using Chinese data, changing the output layer to enable multi-class classification.

After observing the mean, median, and standard deviation of the number of labels in the training data, we adjust our label selection criteria to mimic that distribution of labels. For this multi-label classification problem, we apply the sigmoid function on the raw output and use the resulting probabilities to select labels. We found that selecting up to 10 labels with a probability larger than 0.2 yields a similar distribution of labels to that in the training data so that the mean, median, and standard deviation are roughly comparable between the training data and our model outputs. For all Chinese models, we discard those articles with empty ESG label fields[5].

4.2 English and French

Given the small size of the individual English and French datasets, we finetune bert-base-multilingual-cased on the combined datasets for single-label classification. We concatenate the article title and content separated by "||" and use it as input for the model. 80% of the combined dataset with 2,399 French and English articles is used for training, and the rest is used for validation. We perform 5-fold cross-validation and use a majority vote from the five predictions to choose the final

[2] https://www.sasb.org/standards/materiality-finder/?lang=en-us

[3] https://platform.openai.com/docs/models/gpt-3-5

[4] https://huggingface.co/bert-base-multilingual-cased

[5] As there was no specific [not ESG-related] label provided and some non-labeled articles appeared related to ESG issues, we believed these datapoints represented noise.

ESG label. The learning rate is set at $2.5e-5$ and the model is trained for 10 or 20 epochs.

We also finetune multilingual BERT for single-label classification by augmenting the training data using synthetic data described in subsection 3.1 and/or using the given Chinese dataset. We validate these models using the given English and French articles. We train different models using three levels of augmentation - (1) use synthetic data in English and French, (2) use the given Chinese dataset for augmentation, and (3) use synthetic articles in all three languages along with the given Chinese dataset. At inference time, we generate our predictions on the test set using a majority voting ensemble from the models trained on each fold.

T5-based Filtration In converting the ESG labels to indices on a probability vector, we ignore features embedded in the labels' text. For example, the semantic distance between "Toxic Emissions & Waste" and "Packaging Material & Waste" is arguably smaller than between "Board" and "Pay". In the traditional multiclass classification paradigm, these relationships are ignored.

To remedy this, we experiment with using variants of T5 for multiclass classification. Specifically, we take those labels from the French and English data with the top-k highest softmax probabilities as judged by the classifier described in Section 4.2 and encode them with T5 alongside the article content and title. The T5 decoder selects one of the top-k labels as a final prediction. Setting k too high results in too large of a search space, sometimes resulting in context overflow. Setting k too low can cause the fatal mistake of the gold label being absent in the input to T5, dooming the filtration model to an incorrect prediction. We experiment with two multilingual T5 variants: ByT5 (Xue et al., 2022) and MT5 (Xue et al., 2021).

5 Results

5.1 Chinese

Due to the reasons stated in Section 4.1, we exclude those non-labeled data points. Thus, the reported model results do not consider those articles. Our Chinese model achieves the following results on the Chinese validation dataset obtained from the training data provided: **F1 score: 36.81, precision: 29.10, and recall: 52.62**. The three versions of our submissions to the official test set are the results from the single model with the above performance using different label selection criteria.

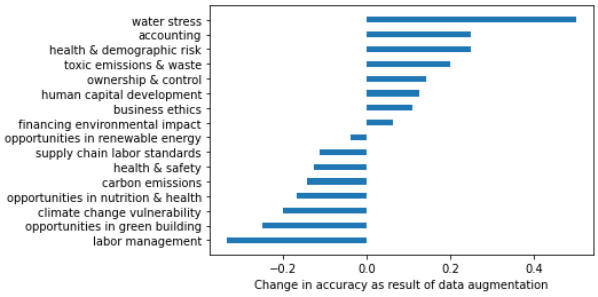

Figure 2: Classes with change in accuracy in the French test set due to augmentation using synthetic English and French data on training for 20 epochs.

5.2 English and French

Table 1 shows the precision (P), recall (R), and F1 score for English and French articles in the test set. Figure 2 shows the change in class-wise performance when using the synthetic counterparts.

Best French Result Finetuning multilingual BERT for 20 epochs on the original English and French training dataset augmented with synthetic articles in English and French achieves the F1 score of 79.96 for the French articles in the test split. Using data augmentation increases the F1 score by 1.39 when trained for 10 epochs. When trained for 20 epochs, using data augmentation increases the F1 score by up to 0.63.

Best English Result Finetuning multilingual BERT for 20 epochs on the original English and French training dataset augmented with the original Chinese dataset achieves the best F1 score of 66.33 for the English test split. As shown in Table 1, the augmentation increases the F1 score by a maximum of 0.66.

5.3 Classification with T5

As shown in Table 2, the multilingual variants of T5 were not successful in filtering the top-k predictions of the original BERT-based classifier. In the best setting, an mt5-large[6] model was able to boost the English F1 score by +0.77 when provided with the ranked top-5 predictions of the BERT classifier.

Analyzing the outputs of the mt5-large model, it suffered from a strong tendency to hyper-fixate on the positional signal provided by the ranked inputs. Specifically, the mt5-large model only predicted a label different from what the original classifier

[6] https://huggingface.co/google/mt5-large

Training Data	Epochs	$\mathbf{P}_{en}$	$\mathbf{R}_{en}$	$\mathbf{F1}_{en}$	$\mathbf{P}_{fr}$	$\mathbf{R}_{fr}$	$\mathbf{F1}_{fr}$
en + fr (**Jetsons_3**)	10	65.36	66.00	64.88	78.39	78.33	77.38
en + fr	20	66.77	66.67	66.01	80.00	80.00	79.33
en + fr + Syn (**Jetsons_2**)	10	64.11	65.00	63.75	80.34	79.00	78.77
en + fr + Syn	20	64.31	64.67	63.90	**81.32**	**80.33**	**79.96**
en + fr + zh	20	**66.96**	**67.33**	**66.33**	81.05	79.67	79.68
en + fr + zh + Syn_{all}	20	66.63	66.33	65.45	80.34	79.67	79.25

Table 1: Results of finetuning multilingual BERT on the English(en) and French(fr) articles in the test set with and without data augmentation. Official submissions on the test set are designated in bold. Syn - Synthetically generated en and fr articles, Syn_{all} - Synthetically generated en, fr, and zh articles.

	Model	K	EN F1	EN F1 Change	FR F1	FR F1 Change
Labels Shuffled	byt5-base	5	28.08	-28.39	32.08	-42.10
	mt5-base	10	25.58	-30.89	30.37	-43.91
	mt5-large	5	33.89	-22.58	39.1	-35.10
Labels Ranked by Logits	byt5-base	10	56.63	+0.16	74.18	+0.0
	mt5-large	5	**57.24**	**+0.77**	73.74	-0.44

Table 2: Results of the various T5-based models for filtering the top-k predictions made by the initial BERT-based classifier. We use the predictions from the 1st fold of **Jetsons_2** in these experiments.

predicted in 6 out of 600 instances, resulting in a 0.77 improvement in English samples. The model appears to get stuck in a local minimum in that merely predicting the label that appears first gives decent performance (whatever the original BERT-based classifier achieved). In an attempt to solve this hyper-fixation on positional signals, we run experiments with shuffled label inputs as well. This further highlighted the inability of the T5 variants to perform well in this task.

6 Analysis

6.1 Synthetic Data vs. Original

Surprisingly, the synthetic data generated using the methods described in Section 3.1 did not always improve performance on the final test set. To explore this further, we plot the embedding representation of the synthetic and original training data in Figure 3. Embeddings were generated using the paraphrase-MiniLM-L6-v2 model (Reimers and Gurevych, 2019), and cast to a 2-dimensional space using TSNE (Van der Maaten and Hinton, 2008). Qualitatively, we see that the synthetic data seems to have a lower variance in this embedding space than the original data.

Many of the synthetic data points appear to be simpler to classify than the original data points. Notably within the EN + FR data, 185 synthetic

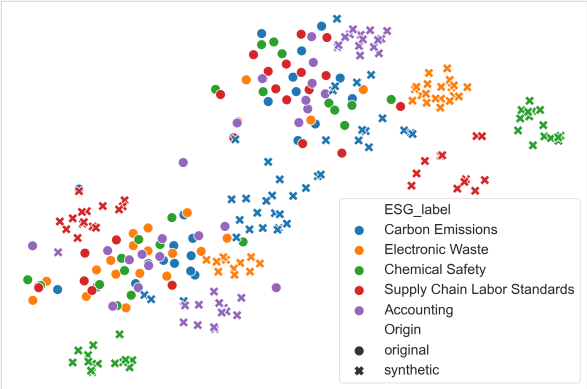

Figure 3: Plotting the embedding space of the original EN+FR articles against the synthetic articles generated with the LLM described in Section 3.1. The synthetic data points appear to be more tightly grouped together than the original training data.

article titles contained at least one token appearing in the gold label, whereas only 6 of the original data points contained this token overlap[7]. This represents a token overlap rate of 62.29% for the synthetic data and only 2.79% for the original data.

7 Conclusion

Excelling at the task of ESG issue identification moves the field of financial NLP to a more well-rounded state, where the primary focus of monetary factors is balanced with other qualitative, social factors. We carry out experiments on the FinNLP shared task of fine-grained ESG issue identification, and find that a BERT-based classifier augmented with synthetic data performs best on French and English data. Additionally, we see that utilizing transfer learning boosts performance on Chinese data.

[7]For example, the synthetic article with the title "BP announces net-zero emissions target by 2050" contains a token overlap ("emissions") with the gold label "Carbon Emissions".

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Allen H Huang, Hui Wang, and Yi Yang. 2022. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mDAPT: Multilingual domain adaptive pretraining in a single model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, Hengkui Wu, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

Tim Nugent, Nicole Stelea, and Jochen L. Leidner. 2021. Detecting environmental, social andnbsp;governance (esg) topics using domain-specific language models andnbsp;data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings*, page 157–169, Berlin, Heidelberg. Springer-Verlag.

Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, 2(4):453–468.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Appendix

A.1 Few-Shot Prompt for Data Generation

```
ESG label: Accounting

Type: few-shot prompt for english language

Prompt: Give 10 examples of news related
to ESG (Environmental, Social, Governance)
topic 'Accounting'. Each example should have
a news title, news summary, Tags related
to the article. Generate these examples in
english language. An example news for
'Accounting' topic - News Title:
```

"Institutional Investors: Fossil Fuel Companies Must Align Numbers with Paris Agreement", News Summary: "The group of investors engaged bp in its campaign in November, and sent parallel letters to other energy majors, including Royal Dutch Shell and Total. While Shell and Total have also lowered their oil and gas price assumptions, the statement notes that most other companies have yet to bring their assumptions in line with the Paris Agreement, and calls on them to do so: Other fossil fuel dependent companies need to sit up and take note. Based on our analysis of major European fossil fuel companies 2019 financial statements... it is clear that almost none have aligned their critical accounting assumptions with the Paris Agreement. The level of BP's impairments demonstrates the potential materiality of this risk hidden in companies balance sheets."

Sample Generated Content:

French language -

Titre de l'article: Les entreprises doivent inclure la diversité dans leur comptabilité, selon une étude
Résumé de l'article: Selon une étude du cabinet de conseil PwC, les entreprises doivent inclure la diversité dans leur comptabilité. Cette étude a examiné les pratiques de comptabilité de 100 grandes entreprises aux États-Unis. Les résultats ont montré que seulement 10 informations sur la diversité dans leurs rapports financiers. Les experts affirment que l'inclusion de la diversité dans la comptabilité peut aider les entreprises à mieux comprendre leur personnel et à améliorer la prise de décision.
Tags : Accounting, Diversity, États-Unis.

English language -

News Title: The Impact of Diversity and Inclusion on Financial Performance
News Summary: Diversity and inclusion are increasingly seen as key drivers of financial performance, with studies showing that diverse companies outperform their peers. This article discusses the accounting implications of diversity and inclusion and the need for better diversity reporting standards.
Tags: ESG, diversity and inclusion, financial performance, accounting implications, reporting standards.

A.2 T5 Modeling

Hyperparameter settings All models are fine-tuned with Adafactor (Shazeer and Stern, 2018) and a learning rate of 1e-3, and a batch size of 32. A linear warm-up for the first 10% of training steps is employed, followed by cosine decay.

Inputs We display an example input to the T5 model described in Section 2 below, with the placeholder `{article_content}`.

```
Lenovo, Kuehne+Nagel Partner
on Solution to Offset
Shipping Emissions
with SAF Purchases
| {article_content}
|| LABELS:
opportunities in renewable energy,
carbon emissions,
financing environmental impact,
opportunities in clean tech,
opportunities in green building
```

In the example above, the BERT-based classifier predicted "Opportunities in Renewable Energy". However, the gold label is "Carbon Emissions". By passing in the ranked predictions from the BERT-based classifier, the T5 model is tasked with remedying the mistaken prediction and instead choosing the 2nd highest ranked ESG label.

Effect of k Figure 4 plots the relationship between k and the percentage of ranked data points which would contain the gold label.

A.3 Effects of Data Augmentation

Figures 6, 7, 5, 8 and 9 show change in class-wise accuracy in French and English test set as a result of different settings of data augmentation. For the French articles, performance for classes like Raw Material Sourcing, Labor Management, Opportunities in Green Building, Consumer Financial Protection, Community Relations, and Supply Chain Labor Standards improve on training for 10 epochs using augmented data.

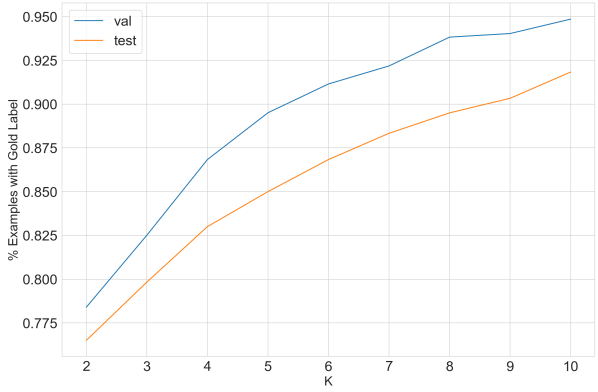

Figure 4: Plotting the relationship between k and the hypothetical upper-bound performance of the T5 model. By setting $k = 5$, 85% of the predictions we pass to T5 includes the gold label on the test split. This represents an upper performance bound of 87% for French and 83% for English.

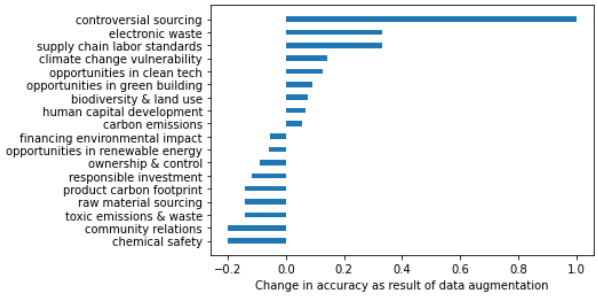

Figure 5: Change in the accuracy of English test instances on training for 10 epochs after augmentation using synthetic English and French data.

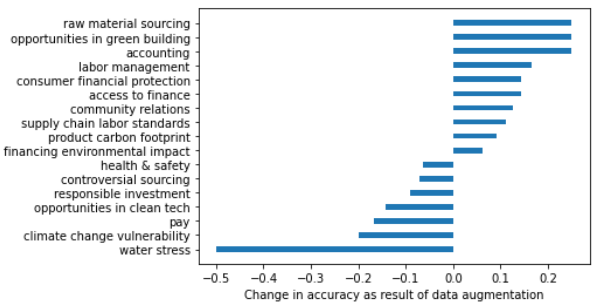

Figure 6: Change in accuracy of French test instances on training for 10 epochs after augmentation using synthetic English and French data.

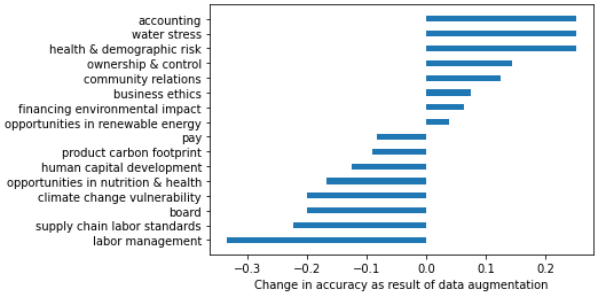

Figure 7: Change in accuracy of French test instances on training for 20 epochs after augmentation using original Chinese dataset.

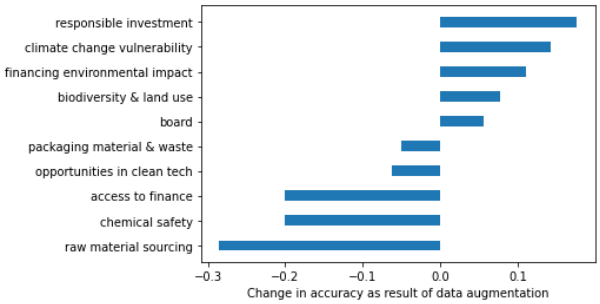

Figure 8: Classes with change in accuracy of English test cases due to augmentation using original Chinese dataset on training for 20 epochs.

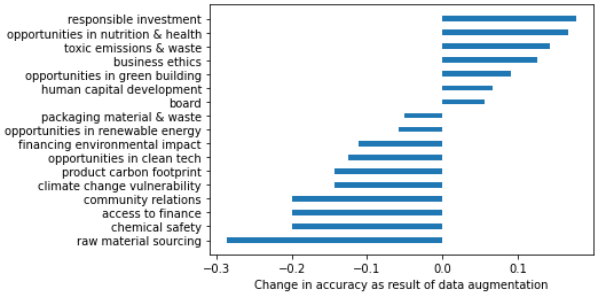

Figure 9: Change in accuracy of French test instances on training for 20 epochs after augmentation synthetic English and French data.