Jetsons at FinNLP 2024: Towards Understanding the ESG Impact of a News Article using Transformer-based Models

**Parag Pravin Dakle, Alolika Gon, Sihan Zha, Liang Wang,
SaiKrishna Rallabandi**, and **Preethi Raghavan**

Fidelity Investments, AI Center of Excellence
{firstname.lastname}@fmr.com

Abstract

In this paper, we describe the different approaches explored by the Jetsons team for the Multi-Lingual ESG Impact Duration Inference (ML-ESG-3) shared task. The shared task focuses on predicting the duration and type of the ESG impact of a news article. The shared task dataset consists of 2,059 news titles and articles in English, French, Korean, and Japanese languages. For the impact duration classification task, we fine-tuned XLM-RoBERTa with a custom fine-tuning strategy and using self-training and DeBERTa-v3 using only English translations. These models individually ranked first on the leaderboard for Korean and Japanese and in an ensemble for the English language, respectively. For the impact type classification task, our XLM-RoBERTa model fine-tuned using a custom fine-tuning strategy ranked first for the English language.

Keywords: ESG, Langauge Models, self-training, multi-lingual

1. Introduction

ESG (environment, social, and governance) related news can impact the performance and reputation of companies, investors, and regulators. One of the key challenges in ESG impact assessment is to estimate the duration of the ESG impact of a news article (Tseng et al., 2023). Different news articles may have different levels of salience, credibility, and relevance for different stakeholders and thus may have different effects on their behavior and outcomes. The LREC-COLING shared task (Chen et al., 2024) presents a multi-lingual impact duration and level classification task based on news articles.

We approach the shared task using the following strategies - (1) Traditional NLP techniques like TF-IDF with logistic regression, SVM (Cortes and Vapnik, 1995), and Random Forest classifiers, (2) De-noising the data to evaluate the impact of removing noisy or less informative samples, (3) Fine-tuning multilingual BERT-style models on individual language and entire dataset, (4) Complementing direct fine-tuning for impact duration with self-training using additional English and French ESG articles, (5) Translating all articles to English to simplify the impact duration task, and (6) Creating an ensemble of the best models for the impact duration task.

2. Related Work

There has been an increased focus on evaluating the nonfinancial activities of a company, which is typically encapsulated under the title of ESG. Park et al. (2022) show that the various topics included in ESG have gradually evolved. Mandas et al. (2023) perform a similar analysis across 11 sectors and show that the best ESG-performing financial institutions are actively committed to the code of best practices in governance. Language Modeling and NLP techniques have been the de facto approaches toward automating the estimation of ESG ratings.

Embeddings for ESG Classification: Raman et al. (2020) investigated employing embeddings from pre-trained language models for classifying sentences relevant to the ESG domain. Mehra et al. (2022) pre-trained a BERT model on ESG-related text, demonstrating improvements in classification tasks related to ESG factors.

Fine-tuning: Nugent et al. (2021) fine-tuned an English BERT-style model specifically for ESG document classification. They explored data generation as an augmentation strategy, enhancing model performance. (Jørgensen et al., 2021, 2023) extended the concept of pre-training language models from financial text to multilingual text and evaluated sentence classification and financial topic classification.

3. Data

The training dataset consists of 2,059 news articles in four languages: 545 English(en), 661 French(fr), 800 Korean(kr), and 53 Japanese(jp) articles. Each article has an associated title and the main content. News articles in all four languages are annotated with impact duration labels: less than 2 years, 2 to 5 years, and more than 5 years[1]. The distribution across the 3 impact duration classes is highly skewed, as shown in Figure 1a. The French and English articles are also annotated with 'low', 'medium',

---

[1]The label names are different for some languages

or 'high' impact level classes (Figure 1c). The Korean dataset also contains impact type annotations with the following classes - opportunity, risk, cannot distinguish. This paper does not focus on the Korean impact type classification task.

Additionally, 31 and 24 duplicates were encountered while pre-processing the data for the Korean and French training data, respectively. We ignore all duplicates with the same class labels, but for 17 of the 24 French duplicates, we randomly select one of the duplicates to be part of the training dataset. This data is split into 10 parts for 10-fold cross-validation with consistent data distribution across all folds in the training and validation sets. The training and validation set lengths were about 1800 and 200, respectively. Lastly, we also found that the test set for Korean contains 1 duplicate, and for Japanese, it contains 19 duplicates and 327 samples with no class label.

4. Impact Duration Task

4.1. Traditional NLP Methods

4.1.1. Baseline Model

The small size of the dataset and high frequency of ESG keywords motivated us to evaluate naive TF-IDF classifier models as a traditional NLP baseline.

We consider logistic regression, random forest, and SVM as our baseline models and adopt 10-fold cross-validation for model training and evaluation. To enable the hyperparameters tuning for those baseline models, we further divide the 10-fold training set into train/val with a ratio of 80/20. Wangperawong (2022) show that using a single vocabulary for all languages and subword tokenization greatly improves the classification results. We use SentencePiece[2] for multilingual tokenization. We convert the obtained tokens to lowercase and compute TF-IDF statistics with filters of maximum frequency(0.7). We tune the penalty parameters C for SVM and logistic regression, and *number of trees, maximum depth parameter, and minimum sample of internal nodes* parameters for the Random Forest (RF) model. The averaged statistics in percentage from the 10-fold testing set are reported in Table 1. The RF model does a good job predicting the impact duration with large variation for the Japanese due to the smaller dataset.

4.1.2. Learning with De-noised Labels

Although the impact duration of the ESG news has been cross-validated with agreement statistics across different annotators, it is sometimes challenging to classify an ESG event into less than 2

years, 2 to 5 years, and more than 5 years window. For example, *The new agreements bring Verizon's projected renewable energy capacity to more than 3GW, enough to power more than 707,000 homes for a year and position the company to meet its goal to source or generate renewable energy equivalent to 50% of its total annual electricity consumption by 2025.* This article was annotated to be '2 to 5 years' probably due to the knowledge of the time difference between 2025 and the year of the annotation. However, the text clearly indicates a time window of one year, which could or should be annotated as "less than 2 years". The ground truth label of this event can hence be ambiguous. Brodley and Friedl (1999) demonstrated that direct training based on the "mislabeled" data generates less desirable models than training with less but de-noised data. Following a similar idea in Wang et al. (2023), we explored a data quality model to score each text-label pair.

Using the RF baseline model fine-tuned on TF-IDF tokens, we evaluate on each of the 10-fold testing sets to obtain the confidence of the prediction $\mathbb{P}$ and the label of the prediction $\hat{\mathbb{Y}}$. Then comparing against the annotation from the ground truth labels $\mathbb{Y}$, we compute a quality score $\mathbb{Q}$: $\mathbb{Y} \times \hat{\mathbb{Y}} \times \mathbb{P} \to [-1, 1]$ using $\mathbb{Q}(\mathbb{P}, \mathbb{Y}, \hat{\mathbb{Y}}) = -\mathbb{P}$ if $\mathbb{Y} \neq \hat{\mathbb{Y}}$ and $\mathbb{Q}(\mathbb{P}, \mathbb{Y}, \hat{\mathbb{Y}}) = +\mathbb{P}$ if $\mathbb{Y} = \hat{\mathbb{Y}}$. Hence, a high-quality score $\mathbb{Q}$ would indicate agreement and high confidence between the predicted labels and the actual labels, whereas a low-quality score $\mathbb{Q}$ indicates agreement with high confidence. Computing on each of the 10-fold testing sets, we obtained the quality score $\mathbb{Q}$ for the entire 2,059 observations, based upon which we delete x% of the data that are potentially of low quality/agreement. Through our evaluation, we have found that deleting 10% of the original data provides a decent improvement with the weighted F1 score shown in Table 1. This indicates a certain level of noisiness within the duration labels.

4.2. Modern NLP Methods

All models described in this subsection have been fine-tuned using 10-fold cross-validation, and the metric used for comparison is the average of weighted F1 scores across the folds. For the winning models, the fold model with the highest evaluation F1 score was further fine-tuned on the dataset for 2 additional epochs.

4.2.1. Fine-tuning Language Models (LMs)

We first fine-tune the XLM-RoBERTa (large) model (Liu et al., 2019) using both the title and the main content of the news articles in each language. We also consider the Longformer (large-4096) (Beltagy et al., 2020) model since some articles surpass
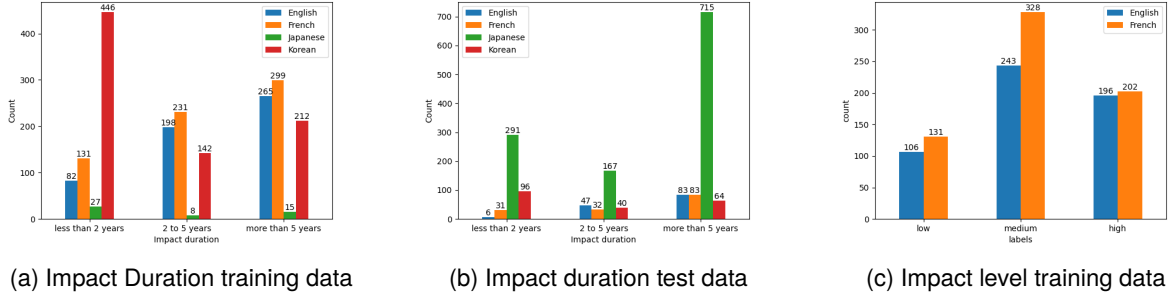
---

[2]https://github.com/google/sentencepiece

| (a) Impact Duration training data | (b) Impact duration test data | (c) Impact level training data |

Figure 1: Data distribution across classes

Model	Setting	overall	kr	jp	en	fr
Logistic	Normal	53.85	63.96	64.78	53.47	50.44
SVM	Normal	53.61	64.02	72.59	51.34	49.27
Random Forest (RF)	Normal	58.23	71.13	61.45	61.45	58.67
Logistic	De-noised	54.61	66.82	72.49	52.36	52.41
SVM	De-noised	55.96	68.42	73.13	54.98	52.79
Random Forest (RF)	De-noised	58.60	71.03	70.00	62.98	60.35

Table 1: Impact Duration: Weighted F1 averaged across 10-folds for baseline models

the maximum token length of conventional BERT-style models(Kannan and Seki, 2023). Given the small size of the datasets per language, specifically Japanese, we also fine-tune multilingual models by combining the news articles in all four languages. Table 2 shows the results of the fine-tuning experiments. We did not fine-tune a monolingual model for Japanese due to the small training data size.

Model-Lang	PSL	en	fr	kr	jp
XLM-en	_	57.9	_	_	_
XLM-fr	_	_	62.6	_	_
XLM-kr	_	_	_	**66.5**	_
LF-en	_	57.7	_	_	_
LF-fr	_	_	**72.9**	_	_
XLM-all	_	57.4	71	66.1	56.7
LF-all	_	50.2	57	39.8	39
XLM-all	direct	**60.2**	71.3	63.4	**59.9**
LF-all	direct	46.6	55.2	44	44.1
LF-all	avg. conf.	50.8	70.4	42.8	58.2

Table 2: Weighted F1 score for impact duration classification averaged across 10-folds for a) fine-tuned LMs (rows 1-7), and b) semi-supervised learning (rows 8-10). Note: **XLM**: XLM-RoBERTa, **LF**: Longformer, **PSL**: Pseudo-label generation methods, **XLM-kr**: Korean_Jetsons_1 submission

4.2.2. Semi-supervised Learning

The training dataset is small and skewed across the impact duration labels. For English and French, 45-48% of the articles belong to the 'more than 5 years' class. For Korean, 55% of the data belongs to the 'less than 2 years' class. To overcome this class imbalance, we use a subset of the news articles released as part of an ESG issues classification task (Chen et al., 2023). We use the XLM and LF models in Table 2 as teacher models and make predictions on the English and French news articles in the ESG issues classification dataset. We generate pseudo labels in two ways: a) **direct**: use the label predicted by the multilingual teacher model directly, b) **avg. conf.**: for each article, take the average of the two confidence scores for each class predicted by the multi-lingual and mono-lingual teacher models and choose the label with the maximum average score. We sample articles based on these pseudo labels and combine them with the original training data to reduce class imbalance. This augmented data is used to fine-tune XLM and LF models. The weighted F1 scores for these models are reported in Table 2. The F1 scores of these models on the final test are reported in section A.2.

4.2.3. English Translation

We also consider converting the problem from multi-lingual to mono-lingual by adding translation as a prerequisite for training and testing. We use the Google Translate API[3] to translate all non-English samples to English. Post translation, we fine-tune a DeBERTa-v3-xsmall model (He et al., 2023) on the class labels using both the article text and title (if available). The model experiment reports a 10-

---

fold average weighted F1 score of 62.37 and a maximum weighted F1 of 66.82 on fold 8. The fold 8 model (DBERT-en) was used in the second submission (<lang>_Jetsons_2) for all languages.

4.2.4. Ensemble

The final model for the impact duration classification task is an ensemble model. We consider an ensemble of the three models - XLM, LF, and the DeBERTa-v3. The class label with the highest total model label score sum is used as the final class label. The submitted ensemble models were - **English_Jetsons_3** (XLM-all-direct, LF-all-avg. conf, DeBERTa-v3), **French_Jetsons_3** (XLM-all, LF-fr, DeBERTa-v3), **Korean_Jetsons_3** (XLM-kr, XLM-all, DeBERTa-v3), and **Japanese_Jetsons_3** (XLM-all-direct, DeBERTa-v3).

5. Impact Level

We conduct experiments with the same two multilingual language models - XLM-RoBERTa-large and Longformer-large-4096 for the impact level task in French and English. We fine-tune the multilingual models in two ways: a) Using both languages, hoping that the data in one language can bolster the performance in the other, and b) separately in each language. First we compare models using only fold 0 data. Table 3 shows that the weighted F1-score for the model trained in combined languages is lower than single language. So we use data in single language to further fine-tune the two models using all 10 folds of data and calculate the average results for each language. It shows that the XLM model has better performance in both languages: XLM-en (**65.02**) vs. LF-en (59.27), and XLM-fr (**65.29**) vs. LF-fr (63.84). We pick the XLM models with the best performance among 10 folds for each language as our first submission. As our second submission, we randomly chose a fold and used the best model fine-tuned on that fold.

| La- | XLM | | | LF | | |
ng	en	fr	all	en	fr	all
en	57.3	_	51.7	56.1	_	47.6
fr	_	71.5	69	_	72.9	59.4

Table 3: Weighted F1 score for impact level classification over the data in the 0th fold

6. Analysis

Table 4 shows the best micro and macro F1 scores on the test set for the submitted models. These models ranked best on 4 out of 7 tasks. Figures 2 and 3 show the confusion matrices for predictions

Submission Model	Micro F1	Macro F1
English_Jetsons_1	64.71	**52.47**
Korean_Jetsons_1	**70**	**66.24**
Japanese_Jetsons_2	36.5	**25.6**
French_Jetsons_1	47.95	37.06
English_Jetsons_1 (IL)	**65.44**	**60.90**

Table 4: F1 scores on the test set, **Bold faced** ones are top on the leaderboard. IL indicates impact level

generated using these four predictions. For impact duration, the models get most confused between 'less than 2 year' and 'More than 5 years' classes. For impact level, 'medium' and 'high' are the most confusing classes.
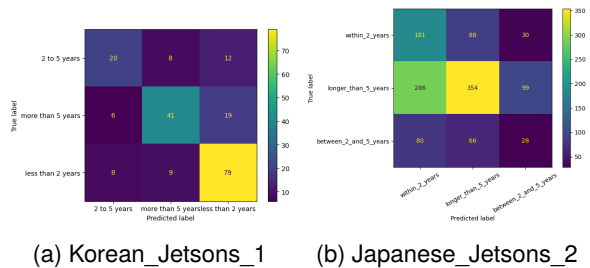

(a) Korean_Jetsons_1   (b) Japanese_Jetsons_2

Figure 2: ID confusion matrices

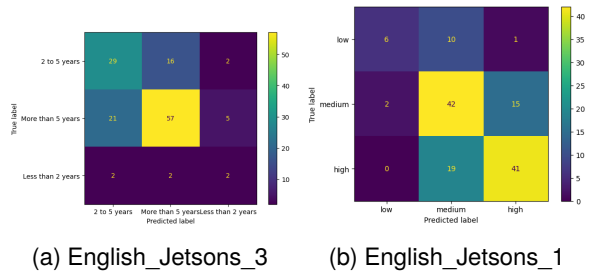(a) English_Jetsons_3   (b) English_Jetsons_1

Figure 3: Confusion matrices for: ID (a) and IL (b).

7. Conclusion

ESG is increasingly important for stakeholders who want to align their values with their investments, reduce risks, and enhance long-term returns. For the FinNLP shared task of impact duration and level classification, we find that finetuning BERT-style models, along with data augmentation techniques like translation and self-training, perform the best. For impact duration in Korean and impact level in English, we find that fine-tuning a BERT-based classifier with a custom strategy performs the best. An ensemble with BERT-style models fine-tuned for impact duration in English using self-training and on just English translations performs best. The DeBERTa-v3 model fine-tuned on only English translations performs best on the Japanese dataset for the impact duration task.

8. Bibliographical References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167.

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)*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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 864–879.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mdapt: Multilingual domain adaptive pretraining in a single model. *arXiv preprint arXiv:2109.06605*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for ESG scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao. -.

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*.

Marco Mandas, Oumaima Lahmar, Luca Piras, and Riccardo De Lisa. 2023. Esg in the financial industry: What matters for rating analysts? *Research in International Business and Finance*, 66:102045.

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 and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

Joonbeom Park, Woojoo Choi, and Sang-Uk Jung. 2022. Exploring trends in environmental, social, and governance themes and their sentimental value over time. *Frontiers in Psychology*, 13:890435.

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.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Liang Wang, Ivano Lauriola, and Alessandro Moschitti. 2023. Accurate training of web-based question answering systems with feedback from ranked users. In *ACL 2023*.

Artit Wangperawong. 2022. Multilingual search with subword tf-idf. *arXiv preprint arXiv:2209.14281*.

A. Additional experiment details

A.1. Traditional NLP methods

The weighted F1 score across different models and different languages is summarized in Figure 5.
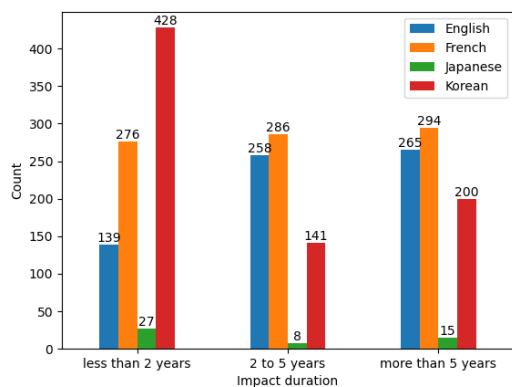
Figure 4: Data distribution across classes in the training set used in semi-supervised learning

A.2. Modern NLP Methods

The distribution of data across class labels in the dataset used to train the student models using semi-supervised learning is shown in figure 4. The micro and macro F1 scores achieved by the different fine-tuned language models on the test set are reported in table 5. In the case of the English, Korean, and Japanese data, we see that the models with the best 10-fold cross-validation scores also perform similarly on the test set. However, for the French news articles, while fine-tuning the Longformer model using only French data (LF-fr) gives maximum average weighted F1 during cross-validation, the same isn't reflected on the test set. XLM-Roberta fine-tuned on articles in all languages along with self-training (row 8 in table 5) gives the best macro F1 of **50.54** and micro F1 of **53.42**. The scores on the Japanese test data have been calculated after removing the 327 unlabelled news articles.

B. Hyperparameters

For the XLM-RoBERTa and Longformer fine-tuning experiments, the learning rates for the mono-lingual and multi-lingual models were $2e-5$ and $8e-6$, respectively, along with batch size -8 and epochs - 10. The Longformer-large models were fine-tuned with gradient accumulation of 2 steps. For the DeBERTa-v3-xsmall model, the following hyperparameters were: learning rate - $2e-05$, epochs - 10, weight decay 0.01, and batch size - 2. The fine-tuning process was carried out on a GPU with 32 GB memory.
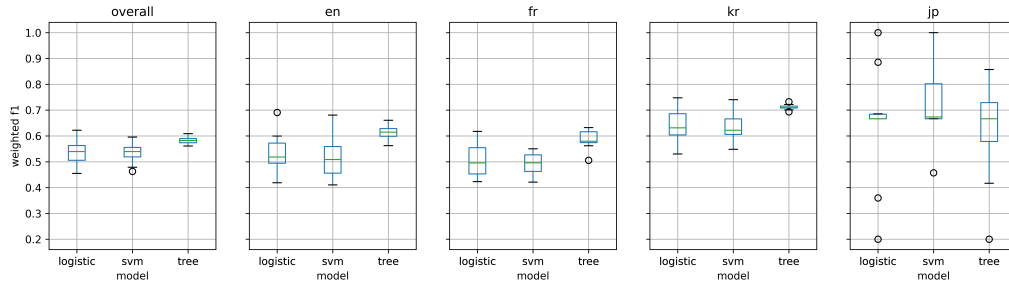
Figure 5: Weighted F1 scores across 10-folds for traditional NLP models for predicting impact length

Model-lang	PSL	en		fr		kr		jp	
		Mi. F1	Ma. F1	Mi. F1	Ma. F1	Mi. F1	Ma. F1	Mi. F1	Ma. F1
XLM-en		**61.03**	**48.79**						
XLM-fr				54.79	43.01				
XLM-kr						**70.0**	**66.24**		
LF-en		55.88	44.9						
LF-fr				47.95	37.06				
XLM-all		58.82	43.66	47.95	43.28	64.0	57.09	**38.19**	**32.99**
LF-all		59.56	38.3	43.84	36.86	48	21.62	24.81	13.25
XLM-all	direct	61.03	46.7	**53.42**	**50.54**	67	62.64	36.23	31.97
LF-all	direct	56.62	47.11	42.47	35.87	50.5	43.62	24.81	13.25
LF-all	avg. conf.	58.82	45.16	38.36	34.34	47.5	27.36	28.3	18.26

Table 5: Micro and Macro F1 for impact length classification task in the final test set.