Classification of multilingual financial articles

Alkhimenkov Leonid

December 2024

Abstract

This paper addresses the problem of classifying multilingual (15 languages) financial article titles. The paper discusses two approaches: using pre-trained multilingual models to learn the classification task and using sequentially trained multilingual translator and classifier models. The work is based on the articles[1]. The link to my project is here: https://github.com/Leonid-Alkhimenkov/FIN_NLP_project.

1. Introduction

The field of financial text is particularly interesting for multilingual NLP, given that it is produced worldwide. The text often includes invoices, transactions, accounting data, tax policies, and stock market information, among others, and there are emerging efforts to create monolingual financial BERTs (Fin-BERTs) for financial text processing (Araci, 2019; DeSola et al., 2019; Yang et al., 2020b; Liu et al., 2021). However, financial text processing by multinational companies is inherently multilingual, so there is a need to develop the most efficient methods to tackle the task. In this paper, this will be done on a dataset that includes article titles in 15 different languages and their category-based classifications. A design approach that can be applied to similar tasks will be developed.

1.1 Team

The project was prepared by Leonid Alkhimenkov.

2. Dataset

The dataset is a financial multilingual corpus consisting of real-world article titles spanning 15 languages, providing two classification tasks: multi-class and multi-label classification.(Table 1).

Example	Lang.	Low-Level labels	High-Level labels
Encuesta Mundial de CEOs 2019 - Hostelería	SPA	· Board, Strategy & Mgmt. · Retail & Consumers	Business & Management
Amendments to VAT legislation	ENG	· VAT & Customs · Government & Policy	Tax & Accounting
Skatta- og lögfræðisvið	ISL	· Tax	Tax & Accounting
Bestyrelsens rolle i forhold til strategiarbejdet	DAN	· Board, Strategy & Mgmt.	Business & Management
Εισαγωγή στην Ελληνική Φορολογία	GRE	· Tax	Tax & Accounting
「事業再編・再生支援」と「ディール戦略」部門を統合・強化	JPN	· M&A & Valuations, · Board, Strategy & Mgmt.	Finance
Veri Analitiği ve Adli Bilişim Çözümleri	TUR	· Financial Crime · Technology	Government & Controls

Table 1: Examples from the MultiFin dataset covering different languages, writing scripts, and combinations of Low-Level and High-Level labels. See Section 3 for more details on the languages and annotation process.

The dataset is based on a collection of public articles published on the websites of a large accounting firm. A portion of the archive was provided for this study. The data collection is based on a real application deployed in a large accounting firm. The choice of language is determined by the branches of the company that provided us with their data. The portion of the archive covers published materials in 15 languages and contains approximately 10,000 headlines.

The distribution of titles by language is shown in Figure 1.

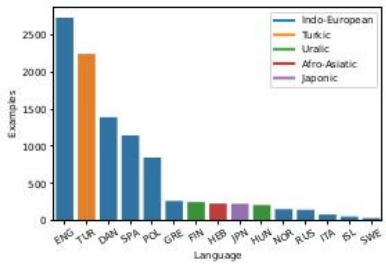

Figure 1: Number of examples per language in MULTIFIN. Bars in the same color indicate these languages belong to the same language family. In this paper, we define languages with more than 500 examples—ENG, TUR, DAN, SPA, POL—high resource languages and the remaining low resource languages.

The publication date is mostly from 2015-2021, and some titles are missing dates. The proposed benchmark contains all languages that have been approved for use, reviewed by experts, ensuring the reliability and quality of both language and content. Although the choice of 15 languages may not be ideal (for example, African and Indian languages are missing, as well as Arabic and Modern Standard Mandarin), I use a massive multilingual dataset for financial NLP.

Annotation scheme The articles were already tagged with predefined topics from the company's internal system. Based on these topics, a new, more general set of labels, called LOW LEVEL, and coarser-grained labels, called HIGH LEVEL, are derived. An overview of LOW-LEVEL and HIGH-LEVEL topics is presented in table 2.

HIGH-LEVEL	LOW-LEVEL
Technology	Technology
	IT Security
Industry	Power, Energy & Renewables
	Supply Chain & Transport
	Healthcare & Pharmaceuticals
	Retail & Consumers
	Real Estate & Construction
	Media & Entertainment
Tax & Accounting	VAT & Customs
	Tax
	Accounting & Assurance
Finance	M&A & Valuations
	Asset & Wealth Management
	Actuary, Pension & Insurance
	Banking & Financial Markets
Government & Controls	Government & Policy
	Financial Crime
	Governance, Controls & Compliance
Business & Management	Board, Strategy & Management
	Start-Up, Innovation & Entrepreneurship
	Corporate Responsibility
	SME & Family Business
	Human Resources

Table 3: Overview of HIGH-LEVEL and LOW-LEVEL topics. The coarse-grained single labels are derived from the fine-grained multi-label annotations based on either a majority-vote, using the first tag in case of ties.

In my solution I will solve the classification problem by LOW-LEVEL topics, covering data with all article titles and articles in English only.

3. Related Work

To evaluate multilingual learning, we train the model on the full training set that contains all 15 languages (called ALL). To evaluate cross-language transfer, we train the model on a subset that contains only the English training data (ENGLISH); and a subst that contains the 5 high-resource languages (i.e., English, Turkish, Danish, Spanish, Polish) (HIGH RESOURCE).

Model Selection In the context of zero-response cross-language transfer, it has been shown that performance on the source language (e.g., English) development set correlates poorly with performance on the target language [2]. They follow the paper [3] and use the joint development set of all languages. Figure 2 is a high-level illustration of our experimental setup. The trained model that achieves the highest Micro F1 score on the development set is ultimately evaluated on the test set. They repeat the experiments five times using different random seeds and report the means and standard deviations.

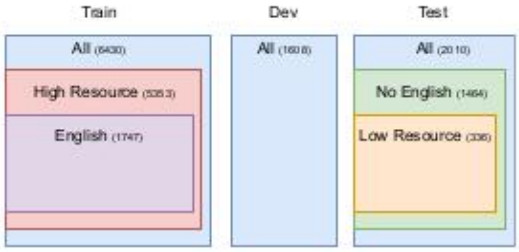

Figure 2: We train models on the complete training set as well as two subsets, to evaluate the multilingual learning and cross-lingual transfer capacities respectively. We use a joint development set of all the languages to select the trained checkpoint. The final model is evaluated on the test and metrics evaluated on the complete test as well as two subsets are reported. Numbers in brackets are the examples belonging to the corresponding (sub)set.

Table 3 shows that models trained on a training set consisting of all languages (ALL) achieve slightly better results (2.0-4.5 absolute F1) than those trained on high-resource languages (HIGH RESOURCE) when the trained models are evaluated on the full test set. However, this performance gap becomes much larger (11.4-30.2 absolute F1) when the models are evaluated on a subset containing only low-resource languages, which is expected since the latter setting requires zero transfer when training on HIGH RESOURCE and evaluating on LOW RESOURCE. In the analysis by languages (detailed in the next section), they also observe that once the training set contains abundant examples (500+) for these languages, the models achieve almost the same results when evaluated on high-resource languages. Therefore, they focus on the evaluation results on low-resource languages.

Model	Training	Test		
		ALL	NO ENGLISH	LOW RESOURCE
FASTTEXT$_{BAG}$	ALL	74.2 $\pm$ 0.2	71.7 $\pm$ 0.2	60.9 $\pm$ 0.8
	ENGLISH	41.8 $\pm$ 1.5	24.5 $\pm$ 1.6	27.9 $\pm$ 3.2
	HIGH RESOURCE	70.3 $\pm$ 1.1	66.8 $\pm$ 1.1	38.2 $\pm$ 1.2
FASTTEXT$_{LSTM}$	ALL	85.4 $\pm$ 0.4	83.6 $\pm$ 0.4	74.4 $\pm$ 0.9
	ENGLISH	51.6 $\pm$ 0.5	36.9 $\pm$ 0.6	41.9 $\pm$ 1.9
	HIGH RESOURCE	82.4 $\pm$ 0.6	80.0 $\pm$ 0.6	59.5 $\pm$ 1.5
sBERT	ALL	73.5 $\pm$ 0.2	67.9 $\pm$ 0.2	52.0 $\pm$ 0.2
	ENGLISH	50.8 $\pm$ 0.5	32.7 $\pm$ 0.4	27.5 $\pm$ 0.6
	HIGH RESOURCE	69.9 $\pm$ 0.3	62.8 $\pm$ 0.5	27.4 $\pm$ 0.2
mBERT	ALL	88.6 $\pm$ 0.3	86.5 $\pm$ 0.3	77.9 $\pm$ 0.5
	ENGLISH	58.3 $\pm$ 0.7	43.5 $\pm$ 1.0	39.4 $\pm$ 2.3
	HIGH RESOURCE	84.1 $\pm$ 0.4	80.6 $\pm$ 0.4	47.7 $\pm$ 0.7
XLM-R	ALL	**90.8** $\pm$ **0.4**	**89.4** $\pm$ **0.4**	**83.9** $\pm$ **0.6**
	ENGLISH	68.0 $\pm$ 1.3	59.2 $\pm$ 1.6	59.8 $\pm$ 1.9
	HIGH RESOURCE	88.6 $\pm$ 0.4	86.4 $\pm$ 0.5	71.0 $\pm$ 1.9
MT5	ALL	81.3 $\pm$ 0.1	76.6 $\pm$ 0.2	51.0 $\pm$ 1.5
	ENGLISH	50.7 $\pm$ 1.0	34.3 $\pm$ 1.1	25.5 $\pm$ 1.9
	HIGH RESOURCE	78.5 $\pm$ 0.3	72.9 $\pm$ 0.5	33.7 $\pm$ 0.2

Table 4: Evaluation results on fine-grained topics (LOW-LEVEL). This is a multi-label classification task with 23 labels, and each example may be assigned up to three topics. All experiments are repeated five times using different random seeds. Averaged Micro F_1 scores and the standard deviations are reported. Best results per column are marked in bold.

The first observation is that different pre-trained multilingual models vary in their ability to learn multilingually on their dataset. That is, when fine-tuned on ALL, the model's performance on low-resource languages ranges from 51.0 to 83.9. The ability to experience zero cross-lingual transfer is another interesting property of multilingual models. Previous studies show that models trained only on English can achieve impressive results on examples in other languages [3].

However, they observe poor performance when the models are trained on ENGLISH and evaluated on LOW RESOURCE (all below 40 F1, except for XLM-R, which achieves around 40 F1). In terms of source language selection, they observe moderate improvements (6.8–11.2 F1) when the multilingual pre-trained models (i.e., MBERT, XLM-R, MT5) are transferred to other languages (HIGH RESOURCE: ENG, TUR, DAN, SPA, POL) rather than just ENGLISH. On the other hand, the improvement becomes much larger (17.6 F1) when FASTTEXTLSTM is trained on more languages, indicating that the model can better leverage information from additional languages than the Transformer-based models. When trained on HIGH RESOURCE,

FASTTEXT LSTM is only slightly inferior to MBERT, and outperforms all other models except XLM-R for the transition from HIGH RESOURCE to LOW RESOURCE. This may be due to the explicit embedding alignment mechanism used in the FASTTEXT approach.

We also calculated the Wilcoxon signed-rank test to assess whether there is a statistically significant difference between the results of XLM-R and MBERT. XLM-R significantly (p-value ≤ 0.05) outperformed MBERT when trained on ALL, ENGLISH, and HIGH RESOURCE, and then evaluated on the full test set. However, the differences for individual languages were not always statistically significant ($p > 0.05$). When both models were trained on ALL, the differences in results on TUR, NOR, RUS, SWE, ITA, and ISL were not significant; the same applies to the difference in ENG when studying in ENGLISH, as well as the difference in SWE and ISL when studying in HIGH RESOURCE.

4. Model Description

At the very beginning of my work, I analyzed the solution from the article and tried to implement it. The problem was solved only for predicting LOW-LEVEL labels. Indeed, the xlm-roberta-base model showed the best result on ALL. Later, I found a model trained on financial data - ProsusAI/finbert, which turned out to be slightly better than xlm-roberta-base having fewer parameters (110 vs. 279). The structure at that time was standard for transformer models. During the study, it turned out that in the case of training and predicting data only in English, the micro F1 score increases significantly (by 0.05) on the models selected as the main ones (xlm-roberta-base and ProsusAI/finbert), so it was decided to use two models sequentially - facebook/m2m100_418M - a multilingual text translator model with a huge number of languages. It will translate text from various languages into English. Next, ProsusAI/finbert (xlm-roberta-base was also tested) trained for the multiple classification task will be used. The block diagram is shown in Figure 3.

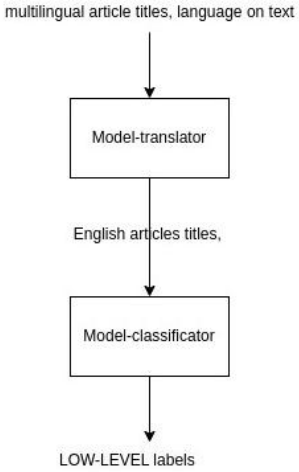

Figure 3: New structure.

5 Experiments.

5.1 Metrics

Micro F1 Score

The **Micro F1 Score** is a metric that balances precision and recall across all classes by treating them as a single group. It is computed based on the total counts of true positives (TP), false positives (FP), and false negatives (FN) across all classes. This metric is particularly useful for datasets with imbalanced classes, as it gives equal importance to all instances.

Formula:

$$\text{Micro F1} = 2 \cdot \frac{\text{Micro Precision} \cdot \text{Micro Recall}}{\text{Micro Precision} + \text{Micro Recall}}$$

Where:

- **Micro Precision** is calculated as:

$$\text{Micro Precision} = \frac{\sum \text{TP}}{\sum(\text{TP} + \text{FP})}$$

- **Micro Recall** is calculated as:

$$\text{Micro Recall} = \frac{\sum \text{TP}}{\sum(\text{TP} + \text{FN})}$$

Variables:

- TP— True Positives (correctly predicted positive instances).
- FP — False Positives (incorrectly predicted positive instances).
- FN — False Negatives (missed positive instances).

Note: In micro-averaging, TP, FP, and FN are aggregated across all classes before computing precision and recall.

5.2 Experiment Setup

In the beginning, there were about 10 runs searching for optimal hyperparameters for the following models: xlm-roberta-base, bert-base-multilingual-cased and google/mt5-base.

Among them, the best was xlm-roberta-base with the following hyperparameters: num_train_epochs=10, learning_rate=2.5e-5, weight_decay=0.02.

Later, there were several runs of the ProsusAI/finbert model. Its best hyperparameters were: num_train_epochs=10, learning_rate=2e-5, weight_decay=0.01.

The runs of xlm-roberta-base and ProsusAI/finbert for the English-only dataset had the same hyperparameters as in the previous experiments. Running the translator model (without hyperparameters) together with the classifier model (xlm-roberta-base and ProsusAI/finbert) had roughly the same optimal hyperparameters.

5.3 Baselines

We will choose the MT5 model as the base model on ALL LOW-LEVEL labels (0.813), since its results are presented in the table and are about average.

6. Results

The results were as follows and are presented in Table 4:

Experiment	Micro F1 score
ALL data and xlm-roberta-base	0.869
ALL data and ProsusAI/finbert	0.877
Only Eng data and xlm-roberta-base	0.918
Only Eng data and ProsusAI/finbert	0.926
ALL data and facebook/m2m100_418M + xlm-roberta-base	0.849
ALL data and facebook/m2m100_418M + ProsusAI/finbert	0.855

Table 4: results (only LOW-LEVEL).

7 Conclusion

As a result of the work, the article [1] was reviewed, based on the dataset of which the model they had previously made was implemented, an experiment was conducted showing that monolingual financial models provide a significant increase in metrics compared to multilingual ones, and an approach was also devised that in the future, with more highly developed translation models, would outperform each other in quality. approaches.

Rerefences

[1] Rasmus Kær Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, Desmond Elliott MULTIFIN: A Dataset for Multilingual Financial NLP // 2023.

[2] Yang Chen, Alan Ritter Model selection for cross-lingual transfer // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5675–5687, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021.

[3] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, Veselin Stoyanov XNLI: Evaluating Cross-lingual Sentence Representations // In EMNLP, 2018.