MultiFin: A Dataset for Multilingual Financial NLP

Rasmus Kær Jørgensen[1,2] **Oliver Brandt**[3] **Mareike Hartmann**[5,6] **Xiang Dai**[4]
Christian Igel[1] **Desmond Elliott**[1]
[1]Department of Computer Science, University of Copenhagen
[2]PricewaterhouseCoopers (PwC) [3]Independent Researcher [4]CSIRO Data61
[5]Department of Language Science and Technology, Saarland University
[6]German Research Center for Artificial Intelligence (DFKI)
rasmuskj,xiang.dai,igel,de@di.ku.dk
obrandt2311@gmail.com mareikeh@lst.de

Abstract

Financial information is generated and distributed across the world, resulting in a vast amount of domain-specific multilingual data. Multilingual models adapted to the financial domain would ease deployment when an organization needs to work with multiple languages on a regular basis. For the development and evaluation of such models, there is a need for multilingual financial language processing datasets. We describe MultiFin– a publicly available financial dataset consisting of real-world article headlines covering 15 languages across different writing systems and language families. The dataset consists of hierarchical label structure providing two classification tasks: multi-label and multi-class. We develop our annotation schema based on a real-world application and annotate our dataset using both 'label by native-speaker' and 'translate-then-label' approaches. The evaluation of several popular multilingual models, e.g., mBERT, XLM-R, and mT5, show that although decent accuracy can be achieved in high-resource languages, there is substantial room for improvement in low-resource languages.

1 Introduction

Natural language processing technology has substantially improved in recent years due to the general-purpose Transformer model (Vaswani et al., 2017), large-scale self-supervised training from unlabelled corpora (Devlin et al., 2019), and the scaling of both of these to increasingly large datasets and models (Raffel et al., 2020). Nevertheless, there are still benefits to having domain-specific models (Gururangan et al., 2020), especially when working with clinical (Dai et al., 2022) or financial text (Araci, 2019).

The domain of financial text is particularly interesting for multilingual NLP, given that it is produced across the world (Lewis et al., 2004; Kær Jørgensen et al., 2021). The text often includes invoices, transactions, accounting data, tax policies, and stock market information, *inter-alia*, and there is an emerging effort to create monolingual financial BERTs (FinBERTs) to process financial text (Araci, 2019; DeSola et al., 2019; Yang et al., 2020b; Liu et al., 2021). However, the handling of financial text by multinational companies is inherently multilingual, therefore, there is is a need for datasets to evaluate how well models can process multilingual financial text.

To this end, we introduce the MultiFin dataset, a publicly available financial dataset consisting of real-world financial article headlines in 15 languages (see examples in Table 1). MultiFin is annotated with High-Level and Low-Level topics for multi-class and multi-label classification, respectively. The dataset is intended as a resource for developing multilingual financial language models. It is the first benchmark for evaluating cross-lingual and multilingual performance of financial models across multiple languages, writing systems and language families that reflects the real-world multilingual situation in the financial domain.

We benchmark four large-scale pretrained language models (SentenceBERT, mBERT, XLM-R, and MT5) and find that the benefits of large-scale pretraining also apply to financial text. XLM-R is clearly the best performing model in all of our experiments, however, there is a subsantial gap in performance between high- and low-resource languages in MultiFin. Moreover, a simple LSTM initialized with FastText word embeddings gives surprisingly competitive performance in several experiments. Overall, we find the financial domain can benefit from multilingual NLP, and future work should focus on domain adaptive efforts and improving models' capacity to generalize to low-resource languages.

Contributions Our contributions are as follows: (a) We present a multilingual financial dataset based on article titles in multiple languages and annotated with two levels of topics. The dataset is made publicly available at https://github.com/RasmusKaer/MultiFin. (b) We evaluate dif-

Findings of the Association for Computational Linguistics: EACL 2023, pages 894–909

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.

ferent multilingual models under different setups in conjunction with analysis on the multilingual MultiFin to establish baselines for the benchmark. (c) Our analysis identifies a need for further research in minimizing the performance gap between high and low-resource languages, and domain adaptive efforts maybe be a promising direction for narrowing this gap.

2 Existing Datasets for Financial NLP

Financial NLP is an emerging area of NLP. Researchers and practitioners have a keen interest in processing natural language for different downstream tasks in the financial domain, such as text mining in accounting (Loughran and Mcdonald, 2016), financial transactions (Jørgensen and Igel, 2021), sentiment analysis (Malo et al., 2014), and text classification (Arslan et al., 2021). Also, financial economics research shows that news articles and media can be used to forecast firm performance (Tetlock et al., 2008), predict stock market volatility (Glasserman and Mamaysky, 2019) and predict market return (Tetlock, 2007). Moreover, Qin and Yang (2019) show that textual transcripts in combination with audio recordings of company earnings conference calls can be used to predict stock price volatility.

There is a large variety of downstream NLP tasks in the financial domain. However, most work within the community is carried out in a monolingual English setting, where the focus is on adapting successful generic monolingual models to the financial domain (Araci, 2019; DeSola et al., 2019; Yang et al., 2020b; Liu et al., 2021). Only a little work on multilingual domain-adapted models has been investigated (Kær Jørgensen et al., 2021). Since the financial environment is indeed multilingual, further progression is conditioned on the availability of multilingual resources to develop new methods for multilingual NLP in the financial domain.

Datasets in the financial domain An extensive literature review identifies the datasets used for financial NLP. We define three criteria for being assigned to the list: (1) the dataset needs to be publicly available and accessible, (2) it needs a clear definition of the task with accompanying annotations (i.e., labels, tags, etc.), and (3) it needs to be peer-reviewed and documented. These criteria are set to ensure the quality of the data resource and proper availability and accessibility. Table 2 presents our findings.

An investigation of the datasets shows that most resources are in English. Table 2 (A) presents an overview of the English evaluation datasets. AnalystTone Dataset (Huang et al., 2014), FinTextSen (Cortis et al., 2017) and Financial Phrase Bank (Malo et al., 2014) are among the most popular datasets. Sentiment analysis is the most frequent task for the datasets, followed by classification. Only few non-English and multilingual datasets exist. Table 2 (B) and (C) shows available datasets in other languages than English. There are five multilingual datasets which contain English plus three additional non-English languages. The dataset containing most languages is the trilingual (El-Haj et al., 2022) and (Gaillat et al., 2018). In addition, we found three low-resource monolingual sentiment datasets: Arabic BORSAH (Alshahrani et al., 2018), Greek FNS-2022 Shared Task (El-Haj et al., 2022) and the Danish DanFinNews (Kær Jørgensen et al., 2021) which is the Danish equivalent to the Financial PhraseBank.

The need for a multilingual financial resource has

(A) Datasets in English		(B) Non-English datasets		lang
AnalystTone Dataset (Huang et al., 2014)	SA	DanFinNews (Kær Jørgensen et al., 2021)	SA	DAN
FinTextSen (Cortis et al., 2017)	SA	CorpusFR (Jabbari et al., 2020)	NER,RE	FRE
Financial Phrase Bank (Malo et al., 2014)	SA	BORSAH (Alshahrani et al., 2018)	SA	ARA
FiQA Dataset (Maia et al., 2018)	SA,QA			
FinNum-1 (Chen et al., 2018)	Numeral CLS	**(C) Multilingual datasets**		
M&A dataset (Yang et al., 2020a)	Deal completeness CLS	ENG-CHI Parallel Fin. Dataset (Turenne et al., 2022)	TC,MT	ENG,CHI
FinNum-2 (Chen et al., 2019a)	Numeral attachment	FNS-2022* Shared Task (El-Haj et al., 2022)	SA	ENG,SPA,GRE
StockSen* (Xing et al., 2020)	SA	SEDAR* (Ghaddar and Langlais, 2020)	MT	ENG,FRE
FinCausal* (Mariko et al., 2020)	RC,RE	FinSBD-2019* (Azzi et al., 2019)	SBD	ENG,FRE
MultiLing2019 (El-Haj, 2019)	Summarization	SIXX-Corpora* (Gaillat et al., 2018)	SA	ENG,SPA,GER
FIN5 & FIN3 (Salinas Alvarado et al., 2015)	NER			
Stock-event (Lee et al., 2014)	Stock Price Prediction	**(D) Our dataset**		
News-sample OMX Helsinki* (Malo et al., 2013)	SA	MULTIFIN (this paper)	TC	ENG,DAN,FIN,GRE,HEB,HUN,ISL,
EarningsCall (Qin and Yang, 2019)	Stock Price Volatility			ITA,JPN,NOR,POL,RUS,SPA,SWE,TUR
Stocknet (Xu and Cohen, 2018)	Stock Movement Prediction			

Table 2: A list of datasets for financial NLP with corresponding task (SA=Sentiment Analysis, NER=Named Entity Recognition, QA=Question Answering, TC=Topic Classification, RC=Relation Classification, RE=Relation Extraction, MT=Machine Translation, SBD=Sentence Boundary Detection, CLS=Classification). Marked (*) refers to datasets were a request is needed or an application for permission needs to be obtained before that dataset is shared.

been highlighted in several studies (Gaillat et al., 2018; Kær Jørgensen et al., 2021; Jabbari et al., 2020) and its lack of multilingual resources is a limitation for further progression. There is also a need for including different language families and low-resources languages into the research landscape to ensure that not only the high-resources languages lays the foundation of research (Alshahrani et al., 2018). This suggests a gap in resources necessary to advance the financial NLP towards a more multilingual scenario that simulate the financial domain's multilingual environment. Our work, see Table 2 (D), is motivated by creating a gold standard for benchmarking financial models to facilitate work on adapting to multiple languages within a specific domain.

3 The MULTIFIN dataset

The MULTIFIN dataset is a multilingual corpus, consisting of real-world article headlines covering 15 languages. We annotate the corpus using hierarchical label structure, providing two classification tasks: multi-class and multi-label classification.

Data collection The dataset builds on a collection of public articles published on a large accounting firm's websites. A subset of the archive was made available for this study. The data collection is based on a real-world application deployed in a large accounting firm. The language selection is determined by the company branches that made their data available to us. We build a multilingual dataset from the headlines of the entire subset that the firm made available. The subset of the archive covers published material in 15 languages and comprises around 10K headlines. The distribution of headlines over languages is shown in Figure 1. The publication date is mainly from the period of 2015 to 2021 with some titles having missing dates. The proposed benchmark contains all the languages we were permitted to use, reviewed by experts, which ensures the reliability and quality of both language and content. While the selection of the 15 languages might not be ideal (e.g., African and Indic languages as well as Arabic and Modern Standard Mandarin are missing), we provide the first massively multilingual dataset for financial NLP, see Table 2 for an overview over currently available datasets. It is also worthy noting that headlines, due to their limited context, poses a great challenge for text classification models deployed in the wild (Chen et al., 2019b). See Figure 6 for the text length distribution across different languages.

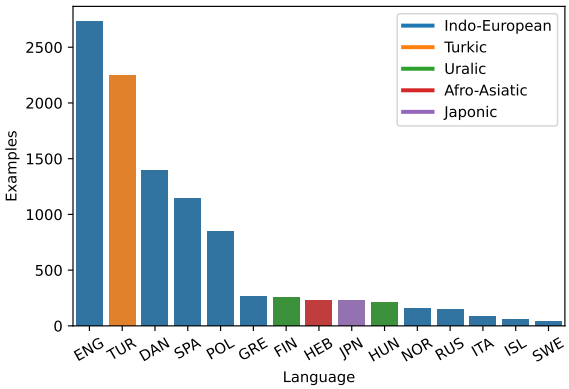

Figure 1: Number of examples per language in MULTIFIN. Bars in the same color indicate these languages belong to the same language family. In this paper, we define languages with more than 500 examples—ENG, TUR, DAN, SPA, POL—high resource languages and the remaining low resource languages.

Annotation Scheme The articles were already tagged with internally pre-defined topics from a company-internal system. Based on these topics, we derive a new, more general label set, referred to Low-Level. Through our label scheme we seek to have different levels of granularity since it gives us the opportunity to go deeper into evaluating the ability of identifying the more refined topics that are presented in titles. Therefore, we first assign fine-grained tags to the topics contain in an headline. For this we use the Low-Level topics. Secondly, we also assign the headline to a single more coarse-grained category, referred to High-Level. We defined the High-Level topics on the basis of universal categories typically found in news media and more common content categorization. Our fine-grained annotation process results in a dataset with multiple labels per headline. We derive High-Level single labels from these multi-label annotations based on either a majority-vote, using the first tag in case of ties. The overview of Low-Level and High-Level topics is presented in 3.

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.

Annotation Process We ask native-level speakers of English and Danish to annotate the dataset using the Low-Level tags. The annotators have domain expertise and participated on a voluntary basis. Detailed annotation guidelines were presented to the annotators before they started. The description contains definitions of topics including some exemplifications of themes and concepts that may occurs for the topics. As for the annotation of multiple labels, the annotators were asked to label up to three topics per example. The annotated labels needed to be ordered by topic weight, i.e., the first annotated topic is the most dominating topic in the sentence, then the second and third most. The overview and statistics of the label distributions can be found in appendix B.

Translate-then-label evaluation We translated the headlines into English for topic annotation using a translation service[1]. We carefully assessed the translation quality to ensure that the translation process does not introduce noise into our dataset. We want to check whether the content of the original sentence is contained in the translation to English. That is, the topics or matters treated in an article stay the same for the translation. For the evaluation, we randomly sample 50 examples from DAN, NOR, ITA, SPA, POL and the entire SWE. We asked evaluators with language proficiency to assess the samples. We presented them with the original sentence, its English translation, and the annotated topics, and ask to answer a true/false question of 1) is the content of the original sentence contained in the English translation, 2) is the property that makes the English sentence fall into this category present in the original sentence as well? The evaluation shows that for DAN, NOR, ITA, SPA, POL and SWE all preserved the properties that make the article fall into a specific category. There was not reported any errors by the evaluators. Thus, we consider translation quality to be high enough to not introduce noise in the process.

Annotator agreement Inter-annotator agreement is measured as multi-label Cohen's κ (Cohen, 1960). The sample selected for evaluation by both annotators is 1200 examples, randomly sampled across languages and topics. The combined κ of 0.94 suggests a near-prefect agreement. Table 5 depicts the topic-level κ.

Description of dataset The dataset consists of 10,048 headlines in 15 languages annotated with 23 topic labels for Low-Level and 6 High-Level topics for multi-class. See Appendix B for details on the distribution of the Low-Level topics and High-Level

[1]Google Translate, version as of Autumn 2021

topics and Appendix E for an overview of the sentence length distribution across different languages. For multi-class, multi-label classification, we have a total of 14,230 tags across 10,048 headlines (80,678 tokens) using 23 fine-grained topics. For multi-class, single label, we have a coarse-grained topic tag for each headline.

4 Experiments and Results

We employ popular pre-trained multilingual models[2] and test their effectiveness under different experimental setups. For experimentation, we will only focus on the LOW-LEVEL multi-label task, and HIGH-LEVEL results are reported in the appendix, Table 9.

4.1 Models

MBERT (Devlin et al., 2019) has been pre-trained on Wikipedia articles of 104 languages. Similarly, **XLM-R** (Conneau et al., 2019) was pre-trained on web crawl data, whose size is much larger than Wikipedia data. For both MBERT and XLM-R, we built a classification layer on top of sentence embedding (i.e., the hidden states corresponding to the first [CLS] token). The classification layer consists of a dense layer and tanh activation function, followed by another dense layer, where the output dimension is the total number of possible topics.

SBERT We use multilingual sentence BERT (Reimers and Gurevych, 2020) to map an input sentence to a 768 dimensional dense vector space and then build a classification layer on top of it. Note that we follow Reimers and Gurevych (2019) to keep the weights of SBERT fixed and use SBERT as a feature extractor. We also investigate the variant of fine-tuning SBERT together with the classification layer. The results of fine-tuning approach are very close to feature extraction approach, although the latter involves much smaller number of trainable parameters (110M vs 600K).

mT5 (Xue et al., 2021) was pre-trained on web crawl data covering 101 languages using a 'text-to-text' format. That is, consecutive spans of input tokens are replaced with a mask token, and then an encoder-decoder transformer is trained to reconstruct the masked-out tokens. When mT5 is used for downstream classification task, the model outputs the literal text of the label instead of a class index.

[2]The number of trainable parameters for each model is listed in Table 8 in the Appendix.

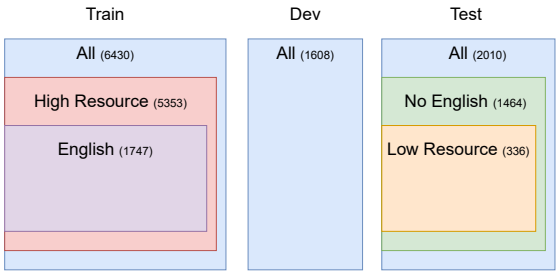

Figure 2: We train models on the complete training set as well as two subsets, to evaluate the multilingual learning and cross-lingual transfer capacities respectively. We use a joint development set of all the languages to select the trained checkpoint. The final model is evaluated on the test and metrics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

In addition to these transformer-based models, we also experiment with models using pre-trained type-based embeddings described below.

Aligned fasttext embeddings As a baseline, we experiment with models using pre-trained type-based embeddings[3], in particular the 300-dimensional fasttext embeddings (Bojanowski et al., 2017) trained on Commoncrawl and Wikipedia data (Grave et al., 2018). In order to enable cross-lingual transfer, we map language-specific fasttext embeddings for all languages covered in our dataset into a common space[4], using RCSLS (Joulin et al., 2018) as a supervised mapping method. Details about embedding alignment can be found in Appendix C. The mapped embeddings are used as inputs for two baseline models: an LSTM classifier ($\text{FASTTEXT}_{\text{LSTM}}$) and a bag-of-embeddings ($\text{FASTTEXT}_{\text{BAG}}$) classifier. The LSTM classifier consists of one bidirectional LSTM layers with a classification layer on top, which receives as input a concatenation of the final hidden states of the top-most layer of forward and backward LSTM. The BoE classifier uses the average over all word embeddings in the input sequence as input to the classification layer. For both models, we use the same classification layer as for the MBERT and XLM-R models.

4.2 Experimental setup

To evaluate multilingual learning, we train the model on the complete training set that contains all 15 languages (referred to as ALL). To evaluate cross-lingual

[3]Fasttext models enable the computation of embeddings for out-of-vocabulary words based on sub-tokens.

[4]We compute pairwise mappings between non-English source embeddings and English target embeddings, and map all non-English embeddings into the space of English embeddings.

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.

transfer, we train the model on (i) a subset that contains only English training data (ENGLISH); and, (ii) a subset that contains 5 high-resource languages (i.e., English, Turkish, Danish, Spanish, Poland) (HIGH RESOURCE).

Model selection In the context of zero-shot cross-lingual transfer, it was shown that performance on a source language (e.g., English) development set does not correlate well with performance in the target language (Keung et al., 2020; Chen and Ritter, 2021). We follow Conneau et al. (2018) and use a joint development set of all the languages. Figure 2 is a high-level illustration of our experimental setup. The trained model which achieves the highest Micro F_1 score on the development set is finally evaluated on the test set. We repeat all experiments five times using different random seeds and mean values and standard deviations are reported.

4.3 Results

Table 4 shows that models trained on the training set consisting of all languages (ALL) achieve slightly better results (2.0-4.5 absolute F_1) than the ones trained on high-resource languages (HIGH RESOURCE) when the trained models are evaluated on the complete test set. However, this performance gap becomes much larger (11.4-30.2 absolute F_1) when models are evaluated on the subset containing only low-resource languages, which is expected, as the latter setting requires zero-shot transfer when training on HIGH RESOURCE and evaluating on LOW RESOURCE. In the per language analysis (detailed in the following section), we also observe that once the training set contains abundant examples (500+) for these languages, models achieve nearly the same results when evaluated on high-resource languages (Figure 3). Therefore, we focus our discussion on the evaluation results on low-resource languages. The first observation is that different pre-trained multilingual models differ in multilingual learning abilities on our dataset. That is, when they are fine-tuned on ALL, model effectiveness on low-resource languages ranges from 51.0 to 83.9 (A detailed analysis can be found in the following section). The ability of zero-shot cross-lingual transfer is another interesting property of multilingual models. Previous studies show that models trained on English only can achieve impressive results on examples in other languages (Conneau et al., 2018; Hu et al., 2020). However, we observe poor performance when models are trained on ENGLISH

and evaluated on Low Resource (all under 40 F_1 except XLM-R achieving near 40 F_1). In terms of the choice of source languages, we observe moderate improvements (6.8-11.2 F_1) when massively multilingual pre-trained models (i.e., mBERT, XLM-R, mT5) are cross-lingual transferred from more languages (High Resource: eng, tur, dan, spa, pol) rather than from English only. On the other hand, the improvement becomes much larger (17.6 F_1) when fasttext$_{\text{LSTM}}$ is trained on more languages, indicating that the model might make better use of information from additional languages than the transformer-based models. When training on High Resource, fasttext$_{\text{LSTM}}$ only slightly underperforms mBERT, and outperforms all other models except XLM-R for transfer from High Resource to Low Resource. This might 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 complete 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 performances on tur, nor, rus, swe, ita, and isl were not significant; the same holds for the difference on eng when trained on English as well as for the differences on swe and isl when trained on High Resource.

5 Analysis and Discussion

Our experiments suggest that although decent accuracy can be achieved for high-resource languages, there is substantial room for improvement in achieving better performance on the multilingual financial dataset. In this section, we present a detailed analysis of the results and investigate some of the findings to identify possible modelling improvements and look into the different dimensions of our dataset.

5.1 Multilingual abilities from a language-level perspective

Multilingual models should ideally learn good representations for all languages they were pre-trained on but this is difficult to achieve in practice due to the "curse of multilinguality" (Conneau et al., 2019). Figure 3 presents per-language results for the three training settings All, English, and High Resource. Generally, we see that XLM-R outperforms the rest of the models across all test settings and languages. When training on All data (first block in Figure 3), although the models have seen all languages during training, mT5 and sBERT seem to be struggling particularly with gre, jpn, heb and hun. We see a drop in performance between high (upper part of the column) and low-resource languages (bottom part of the column), which is expected as the low-resource languages have less examples in the training dataset. When training on High Resource (last block in Figure 3), we observe that performance for the high-resource languages seen during training is stable compared to training on All (indicating that including low-resource languages during fine-tuning does not hurt performance on high-resource languages), but performance for zero-shot transfer to low-resource languages drops significantly. We compare the performance drops suffered on low resource languages from training on All data to training on High Resource data between XLM-R, mBERT, and fasttext$_{\text{LSTM}}$, and find that mBERT suffers from larger performance drops than the other models for most languages, with the largest drops for gre and heb. XLM-R shows the smallest performance drops for most languages, indicating that it has better zero-shot transfer abilities than the other models.

Next, we analyze the best source for zero-shot transfer by comparing the performance on low-resource languages for models trained on High Resource data with models trained on English data. In all cases (except XLM-R on swe), zero-shot transfer works better when more languages are included in the training set. This might be due to the fact that training on more languages allows models to learn more robust representations of input sequences. Another factor might be that, as our dataset has a large label space, including more training examples (regardless of language) can improve learning representations of otherwise sparse classes. As indicated by the averaged results reported in the previous section, for most languages (except fin and isl), fasttext$_{\text{LSTM}}$ shows higher improvements when including more languages to train on.

Comparing zero-shot performance on different target languages for models trained on English (middle block in Figure 3) reveals that all models with a slight exception to XLM-R struggle to generalize to languages not seen during fine-tuning, although they were part of the pre-training languages. Previous research on mBERT suggests a correlation be-

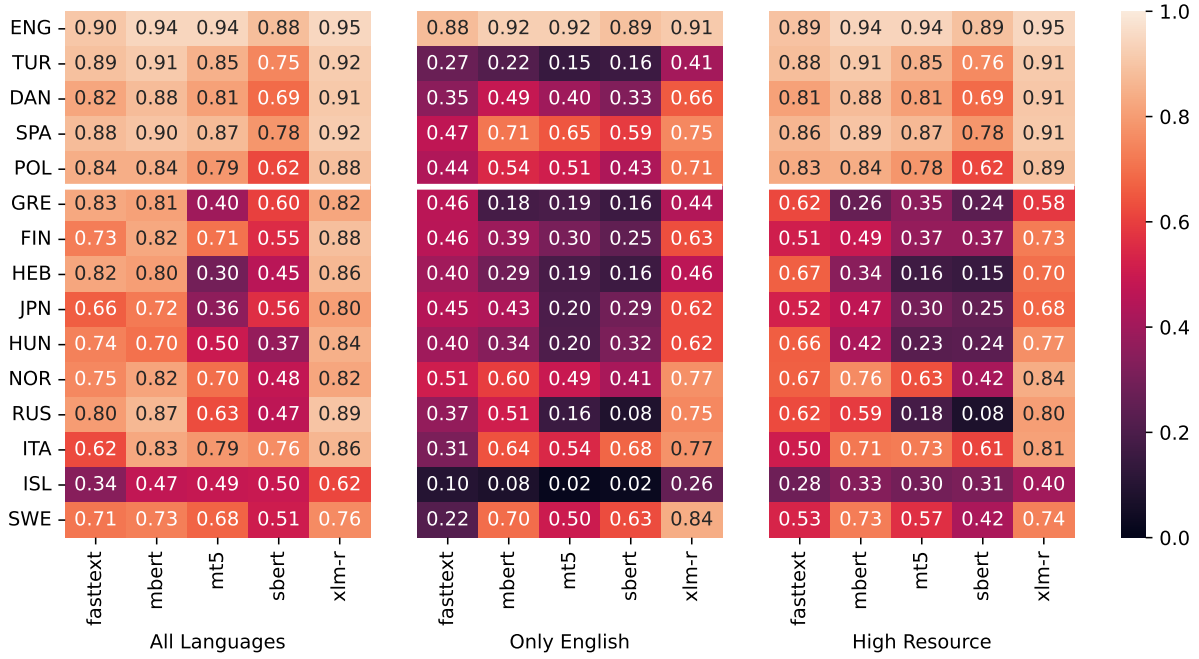

Figure 3: Per language analysis with the multi-label, Low-Level setting. We train on the three settings: All, English, and High Resource and test on All. The first column in each block refers to $\text{FASTTEXT}_{\text{LSTM}}$. Languages are in descending order by the number of examples in MultiFin, with a white separator between high and low-resource languages.

tween zero-shot performance in a downstream task and amount of in-language pre-training data (Wu and Dredze, 2020; Lauscher et al., 2020), which we also observe in our results. Overall, we see very poor generalization ability to certain low-resource languages, such as ISL, GRE, HEB, and RUS. Particularly for ISL, transfer ability from English is nearly non-existing, indicating a need for multilingual models with better transfer abilities to low-resource languages.

Figure 4: The improvement over the vanilla mBERT, in Micro F_1, due to domain-adaptive pre-training mBERT. We compare the model by Kær Jørgensen et al. (2021) against the vanilla mBERT.

5.2 Domain-adaptive pre-training can boost the cross-lingual performance

Domain-adaptive pre-training has been shown to improve the model effectiveness when these models are employed to process domain-specific text (Gururangan et al., 2020). We evaluate the publicly available model by Kær Jørgensen et al. (2021), which continues pre-training mBERT on the combination of multilingual financial text and Wikipedia, and measure the improvement over the vanilla mBERT in Table 4. Note that the multilingual pre-training data in (Kær Jørgensen et al., 2021) cover 9 languages in MultiFin, except POL, GRE, FIN, HEB, HUN, and ISL. Nevertheless, results in Figure 4 show that domain-adaptive pre-trained models outperform vanilla mBERT in all experimental setups, and larger improvements are observed when training set and test set are disjoint, for example, when models are trained on English or high-resource languages and tested on low-resource languages.

5.3 Multilingual versus translate

We assessed that the translation quality was good enough to preserve the topics in Section 3. Therefore, we translate all training and test data to English and fine-tune a monolingual model for English (RoBERTa, Liu et al. (2019)) on the translated training data. We compare performance on the translated

test sets with XLM-R trained and tested on the multilingual data.

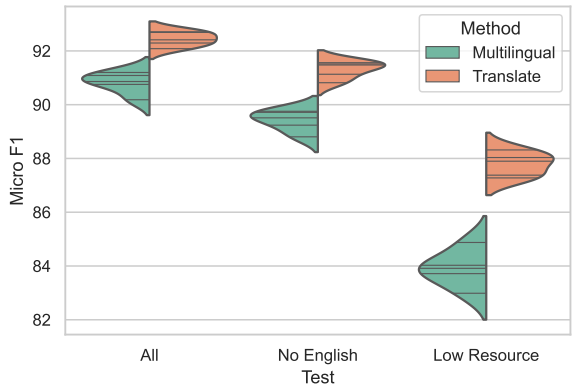

Figure 5: Multilingual (i.e., XLM-R) against translate approach based on English RoBERTa. We use the same setting as in Table 4, where we train on all languages and test on ALL LANG., NOENGLISH and LOWRES.

The monolingual model's advantage of language-specificity over multilingual models (Rust et al., 2021; Rönnqvist et al., 2019) is evident in Figure 5, where the monolingual model trained on English is slightly better than the multilingual model trained on multilingual data.[5] We consider this monolingual model an additional baseline on MULTIFIN.

6 Conclusion

We proposed MULTIFIN, a dataset for the evaluation of multilingual financial NLP models. The main aim is to advance multilingual NLP in the financial domain so it is better suited for new development and evaluation of domain-specific models. MULTIFIN is a diverse dataset with 10,000 examples, covering 15 languages, including different language families and writing systems. We benchmark a collection of standard multilingual language models on MULTIFIN and find that although these models often achieve good performance in high-resource languages, there is a substantial gap in performance between high- and lower-resource languages. The per-language analysis uncovered that most of the benchmarked models do not facilitate a good transfer across the evaluated languages, and for specific languages, indicate a strong need for improving the models' capacity to generalize. The multilingual MDAPT model presented overall better generalization, particularly to low-resource languages, indicating that focusing on multilingual domain-specific methods is a promising direction for future work in financial NLP. Future work includes extending the dataset to include more examples across more languages so better understand the limits of multilingual financial text processing. We are also exploring including the entire document, as opposed to only the headline, but this would depend on high-quality long document processing models (Dai et al., 2022). We hope to motivate and inspire collective work on multilingual NLP in the financial domain.

[5] Artetxe et al. (2020) found that improvements of a translation baseline in a cross-lingual NLI task do not stem from overcoming the cross-lingual gap, but from the fact that translation of the training data introduces alterations which improve generalization to a translated test set. It is possible that in our experiments, the performance of the monolingual model generalizing from translated training data to translated test data is impacted by similar mechanisms.

Limitations

Annotators We are aware that annotators with domain knowledge and language proficiency would be preferred. It was not within our resources to find qualified annotators in the financial domain with expert knowledge and language proficiency for all 15 languages.

Annotation process The number of annotated topics per example is determined to three, although a handful of article titles could potentially be assigned more than three topics. The authors attempted to limit this by prioritizing annotated topics by topic weight (see Section 3).

Acknowledgements

We thank PwC for providing the data and thank Lars Silberg Hansen for his support and valuable contribution to the creation of this dataset.

References

Mohammed Alshahrani, Fuxi Zhu, Mohammed Alghaili, Eshrag Refaee, and Mervat Bamiah. 2018. Borsah: An arabic sentiment financial tweets corpus. In *FNP 2018 —Proceedings of the 1st Financial Narrative Processing Workshop@ LREC*, pages 17–22.

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

Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F. Bissyandé, Jacques Klein, and Anne Goujon. 2021. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 260–268, New York, NY, USA. Association for Computing Machinery.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. 2019. The FinSBD-2019 shared task: Sentence boundary detection in PDF noisy text in the financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 74–80, Macao, China.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019a. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164.

Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 136–143. IEEE.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019b. Deep short text classification with knowledge powered attention. In *AAAI*.

Yang Chen and Alan Ritter. 2021. Model selection for cross-lingual transfer. In *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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. Association for Computational Linguistics (ACL).

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinicio DeSola, Kevin Hanna, and Pri Nonis. 2019. Finbert: Pre-trained model on sec filings for financial natural language tasks. *University of California*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj. 2019. MultiLing 2019: Financial narrative summarisation. In *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, pages 6–10, Varna, Bulgaria. INCOMA Ltd.

Mahmoud El-Haj, Nadhem ZMANDAR, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, Aris Kosmopoulos, Blanca Carbajo-Coronado, and Antonio Moreno-Sandoval. 2022. The financial narrative summarisation shared task (fns 2022). In *Proceedings of the The 4th Financial Narrative Processing Workshop @LREC2022*, pages 52–61, Marseille, France. European Language Resources Association.

Thomas Gaillat, Manel Zarrouk, André Freitas, and Brian Davis. 2018. The SSIX corpora: Three gold standard corpora for sentiment analysis in English, Spanish and German financial microblogs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abbas Ghaddar and Phillippe Langlais. 2020. SEDAR: a large scale French-English financial domain parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3595–3602, Marseille, France. European Language Resources Association.

Paul Glasserman and Harry Mamaysky. 2019. Does unusual news forecast market stress? *Journal of Financial and Quantitative Analysis*, 54:1–38.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *ICML*.

Allen H Huang, Amy Y Zang, and Rong Zheng. 2014. Evidence on the information content of text in analyst reports. *The Accounting Review*, 89(6):2151–2180.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A French corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2293–2299, Marseille, France. European Language Resources Association.

Rasmus Kær Jørgensen and Christian Igel. 2021. Machine learning for financial transaction classification across companies using character-level word embeddings of text fields. *Intelligent Systems in Accounting, Finance and Management*, 28(3):159–172.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. 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.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the importance of text analysis for stock price prediction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1170–1175, Reykjavik, Iceland. European Language Resources Association (ELRA).

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

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*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4513–4519.

Tim Loughran and Bill Mcdonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pyry Takala, Oskar Ahlgren, and Iivari Lappalainen. 2013. Learning the roles of directional expressions and domain concepts in financial news analysis.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Paul C Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Nicolas Turenne, Ziwei Chen, Guitao Fan, Jianlong Li, Yiwen Li, Siyuan Wang, and Jiaqi Zhou. 2022. Mining an english-chinese parallel dataset of financial news. *Journal of Open Humanities Data*, 8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

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.

Linyi Yang, Eoin M Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020a. Generating plausible counterfactual explanations for deep transformers in financial text classification. *arXiv preprint arXiv:2010.12512*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020b. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

A Annotator agreement

The Table 5 below presents the annotator agreement on topic level. The rather high agreement across topics indicate that our annotations are of high quality.

No.	Topic	Kappa, κ
1	Actuary, Pension & Insurance	0.9791
2	Asset & Wealth Management	0.9020
3	Accounting & Assurance	0.9704
4	Banking & Financial Markets	0.9218
5	Board, Strategy & Management	0.9620
6	Power, Energy & Renewables	0.9495
7	Corporate Responsibility	0.9092
8	Media & Entertainment	0.9526
9	Financial Crime	0.9479
10	Government & Policy	0.8889
11	Healthcare & Pharmaceuticals	0.9408
12	Human Resources	0.9537
13	IT Security	0.9346
14	Governance, Controls & Compliance	0.9121
15	M&A & Valuations	0.9617
16	Real Estate & Construction	0.9254
17	Retail & Consumers	0.9526
18	SME & Family Business	0.8670
19	Start-Up, Innovation & Entrepreneurship	0.9888
20	Supply Chain & Transport	0.9321
21	Tax	0.9474
22	Technology	0.9463
23	VAT & Customs	0.9797

Table 5: Full report of inter-annotation agreement of multi-label Cohen's κ.

B Label distribution

We present the distribution of the Low-Level and High-Level topics. In Table 6, we present the distribution over the Low-Level topics. We allowed up-to 3 annotations per examples for the multi-label annotation. This produced a total of 14230 annotation with 1.4 annotations per example on average. In Table 7, we present the distribution over the High-Level topics.

C Cross-lingual transfer with fasttext embeddings

Preprocessing In order to represent inputs with pre-trained fasttext embeddings, we tokenize our data according to how the fasttext training data was tokenized, using Mecab[6] for Japanese, and the tokenizer from the Europarl preprocessing tools[7] (Koehn, 2005) for the other languages.

[6] https://pypi.org/project/mecab-python3/

[7] https://www.statmt.org/europarl/

No.	Topic	Examples
1	Actuary, Pension & Insurance	502
2	Asset & Wealth Management	257
3	Accounting & Assurance	1,452
4	Banking & Financial Markets	782
5	Board, Strategy & Management	866
6	Power, Energy & Renewables	248
7	Corporate Responsibility	277
8	Media & Entertainment	255
9	Financial Crime	310
10	Government & Policy	528
11	Healthcare & Pharmaceuticals	245
12	Human Resources	1,091
13	IT Security	424
14	Governance, Controls & Compliance	501
15	M&A & Valuations	492
16	Real Estate & Construction	351
17	Retail & Consumers	354
18	SME & Family Business	226
19	Start-Up, Innovation & Entrepreneurship	277
20	Supply Chain & Transport	222
21	Tax	1,713
22	Technology	1,169
23	VAT & Customs	1,688
Total		14,230

Table 6: Overview of Low-Level tags across the 23 topics. These represent the 23 labels used in the multi-label task.

No.	Topic	Examples
1	Technology	1,088
2	Industry	1,239
3	Tax & Accounting	3,371
4	Finance	1,447
5	Government & Controls	912
6	Business & Management	1,991
Total		10,048

Table 7: Overview of High-Level tags across the 6 classes. These represents the 6 classes used in the multi-class classification task.

Embedding alignment We map monolingual fasttext embeddings trained on Wikipedia and Commoncrawl into a shared space using RCSLS, by computing pairwise mappings between source languages and English as a target language. As supervision, we rely on the training dictionaries of the MUSE dataset (Conneau et al., 2017), except for Icelandic which is not covered there. For Icelandic, we follow Vulić et al. (2019) in deriving a dictionary based on the Panlex database (Kamholz et al., 2014): We retrieve translations for the 5000 most frequent Icelandic words derived from Opensubtitles published on Wiktionary[8] We only keep single-word transla-

[8] https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Icelandic_

Model	Learning rate	# train epochs	# Params.
FASTTEXT$_{BAG}$	[1e-3,2.5e-3,5e-3,7.5e-3,1e-2,2.5e-2,5e-2]	50	0.1M
FASTTEXT$_{LSTM}$	[1e-3,2.5e-3,5e-3,7.5e-3,1e-2,2.5e-2,5e-2]	50	1.8M/1M/1M
sBERT	[1e-2, 3e-2, 1e-1]	[10, 30, 100]	0.6M
mBERT	[1e-5, 2e-5, 5e-5, 1e-4]	[10, 30, 100]	180M
XLM-R	[1e-5, 2e-5, 5e-5, 1e-4]	[10, 30, 100]	270M
mT5	[1e-4, 3e-4, 1e-3]	[10, 30]	300M

Table 8: The search space of two hyperparameters (learning rate and number of training epochs), as well as the number of trainable parameters for each model. The size of the hidden states in FASTTEXT$_{LSTM}$ is treated as an additional hyperparameter selected from [100,200,300,400,500], hence we report numbers of parameters for three different selected models trained on ALL/ENGLISH/HIGH RESOURCE, corresponding to models with hidden dimensionality 300/200/200, respectively. For all models, we do early stopping on the validation set with a patience of 5 and 10 for transformer-based and fasttext-based models, respectively.

tions. As not all source words are present in Panlex, our final dictionary contains translations for 1,823 Icelandic words. With these dictionaries as supervision, we run RCSLS with default parameters for 10 epochs, and select the best mapping based on the unsupervised selection criterion.

D Experimental Details

For each experiment, we perform grid search to find the best combination of two hyperparameters—number of training epochs and learning rates—on the development set. Table 8 shows the search space of these two hyperparameters as well as the trainable parameters per model.

The particular versions of pre-trained multilingual models can be found at:

- sBERT: https://huggingface.co/sentence-transformers/all-mpnet-base-v2
- mBERT: https://huggingface.co/bert-base-multilingual-cased
- XLM-R: https://huggingface.co/xlm-roberta-base
- mT5 https://huggingface.co/google/mt5-base

Pre-trained fasttext embeddings can be found at:

- https://fasttext.cc/docs/en/crawl-vectors.html

wordlist

E Results of Multi-class classification on HIGH-LEVEL topics

Table 9 show the evaluation results on coarse-grained categories (HIGH-LEVEL), framed as a multi-class classification problem.

F Sentence length distribution

Figure 6 shows the sentence length distribution across languages in the MULTIFIN dataset.

Model	Training	Test		
		ALL	NO ENGLISH	LOW RESOURCE
FASTTEXT$_{BAG}$	ALL	78.1 ± 0.2	76.7 ± 0.8	70.5 ± 1.4
	ENGLISH	60.0 ± 1.0	52.2± 1.1	47.7± 1.1
	HIGH RESOURCE	73.6± 2.4	71.4 ± 2.1	52.8 ± 1.8
FASTTEXT$_{LSTM}$	ALL	83.1 ± 0.7	81.3 ± 0.8	75.9 ± 1.2
	ENGLISH	64.1± 1.5	55.7 ± 1.9	51.6 ± 2.1
	HIGH RESOURCE	80.4 ± 0.4	77.6± 0.5	60.5± 1.5
sBERT	ALL	72.4 ± 0.8	66.1 ± 1.0	55.3 ± 1.8
	ENGLISH	51.9 ± 0.5	38.4 ± 0.8	32.3 ± 0.8
	HIGH RESOURCE	72.1 ± 0.6	65.3 ± 0.7	33.0 ± 1.5
mBERT	ALL	87.4 ± 0.4	85.0 ± 0.4	79.1 ± 0.9
	ENGLISH	60.4 ± 2.4	48.4 ± 3.2	48.1 ± 2.2
	HIGH RESOURCE	82.9 ± 0.5	79.0 ± 0.7	52.3 ± 2.0
XLM-R	ALL	**89.5 ± 0.4**	**87.8 ± 0.5**	**84.0 ± 0.9**
	ENGLISH	74.9 ± 2.2	68.5 ± 2.7	67.9 ± 1.0
	HIGH RESOURCE	87.5 ± 0.7	85.3 ± 0.8	74.7 ± 1.0
mT5	ALL	83.6 ± 0.4	79.7 ± 0.5	61.3 ± 1.2
	ENGLISH	56.6 ± 0.7	42.9 ± 0.8	41.5 ± 1.3
	HIGH RESOURCE	81.1 ± 0.0	76.2 ± 0.1	43.9 ± 0.1

Table 9: Evaluation results on coarse-grained categories (HIGH-LEVEL). Results are averaged over five runs and reported by F1 micro. Multi-class classification task with 6 classes, one per example.

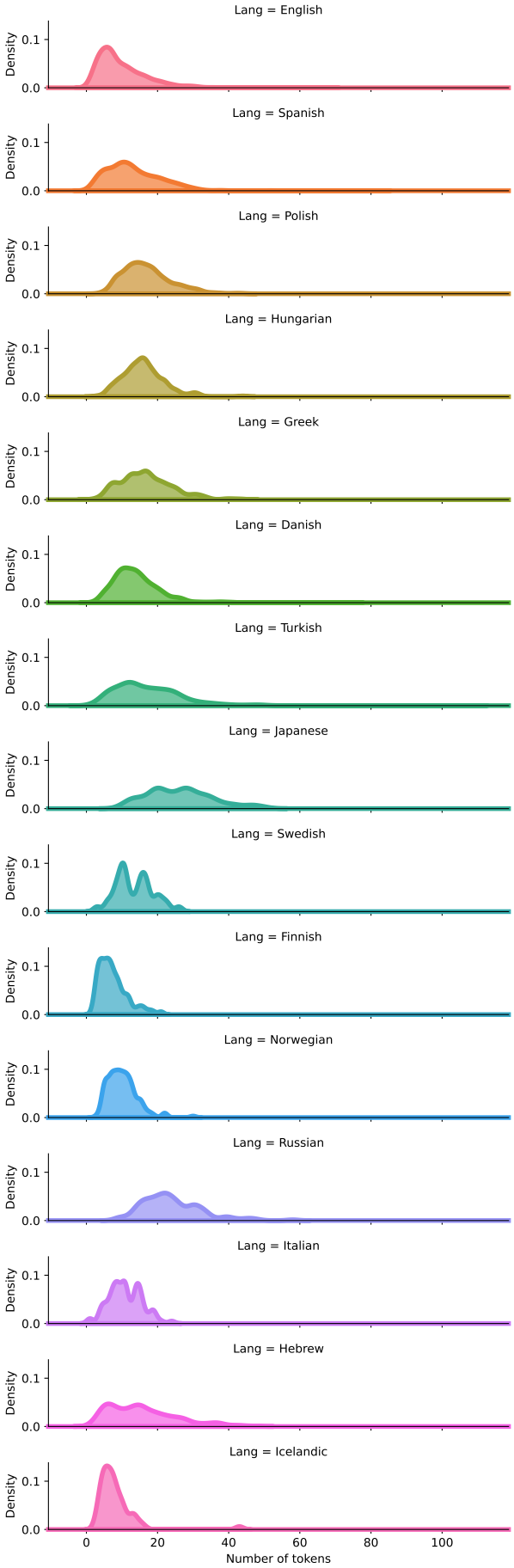

Figure 6: Sentence length distribution across different languages.