MULTIFIN
A Dataset for Multilingual Financial NLP
Jørgensen, Rasmus Kær; Brandt, Oliver; Hartmann, Mareike; Dai, Xiang; Igel, Christian; Elliott, Desmond

Published in:
EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):


**MULTIFIN: A Dataset for Multilingual Financial NLP**

Rasmus Kær Jørgensen1,2  Oliver Brandt3  Mareike Hartmann5,6  Xiang Dai4  Christian Igel1  Desmond Elliott1

1Department of Computer Science, University of Copenhagen  
2PricewaterhouseCoopers (PwC)  
3Independent Researcher  
4CSIRO Data61  
5Department of Language Science and Technology, Saarland University  
6German 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 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 substantial 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-
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.

<table>
<thead>
<tr>
<th>Example</th>
<th>Lang.</th>
<th>Low-Level labels</th>
<th>High-Level labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encuesta Mundial de CEOs 2019 - Hostelería</td>
<td>SPA</td>
<td>Board, Strategy &amp; Mgmt.</td>
<td>Business &amp; Management</td>
</tr>
<tr>
<td>Amendments to VAT legislation</td>
<td>ENG</td>
<td>VAT &amp; Customs</td>
<td>Tax &amp; Accounting</td>
</tr>
<tr>
<td>Skatta- og lögfræðisvið</td>
<td>ISL</td>
<td>Tax</td>
<td>Tax &amp; Accounting</td>
</tr>
<tr>
<td>Bestyrelsens rolle i forhold til strategiarbejdet</td>
<td>DAN</td>
<td>Board, Strategy &amp; Mgmt.</td>
<td>Business &amp; Management</td>
</tr>
<tr>
<td>「事業再編・再生支援」と「ディール戦略」部門を統合・強化</td>
<td>JPN</td>
<td>M&amp;A &amp; Valuations, Board, Strategy &amp; Mgmt.</td>
<td>Finance</td>
</tr>
<tr>
<td>Veri Analitiği ve Adli Bilişim Çözümleri</td>
<td>TUR</td>
<td>Financial Crime</td>
<td>Government &amp; Controls</td>
</tr>
</tbody>
</table>

Different 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 ExistingDatasets 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 Phrase Bank.

The need for a multilingual financial resource has...
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-resource languages into the research landscape to ensure that not only the high-resource languages lay 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.
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.

<table>
<thead>
<tr>
<th>High-Level</th>
<th>Low-Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>Technology</td>
</tr>
<tr>
<td></td>
<td>IT Security</td>
</tr>
<tr>
<td></td>
<td>Power, Energy &amp; Renewables</td>
</tr>
<tr>
<td></td>
<td>Supply Chain &amp; Transport</td>
</tr>
<tr>
<td></td>
<td>Healthcare &amp; Pharmaceuticals</td>
</tr>
<tr>
<td></td>
<td>Retail &amp; Consumers</td>
</tr>
<tr>
<td></td>
<td>Real Estate &amp; Construction</td>
</tr>
<tr>
<td></td>
<td>Media &amp; Entertainment</td>
</tr>
<tr>
<td>Industry</td>
<td>VAT &amp; Customs</td>
</tr>
<tr>
<td></td>
<td>Tax</td>
</tr>
<tr>
<td></td>
<td>Accounting &amp; Assurance</td>
</tr>
<tr>
<td></td>
<td>M&amp;A &amp; Valuations</td>
</tr>
<tr>
<td></td>
<td>Asset &amp; Wealth Management</td>
</tr>
<tr>
<td></td>
<td>Actuary, Pension &amp; Insurance</td>
</tr>
<tr>
<td></td>
<td>Banking &amp; Financial Markets</td>
</tr>
<tr>
<td>Tax &amp; Accounting</td>
<td>Government &amp; Policy</td>
</tr>
<tr>
<td>Finance</td>
<td>Financial Crime</td>
</tr>
<tr>
<td>Government &amp; Controls</td>
<td>Governance, Controls &amp; Compliance</td>
</tr>
<tr>
<td>Business &amp; Management</td>
<td>Board, Strategy &amp; Management</td>
</tr>
<tr>
<td></td>
<td>Start-Up, Innovation &amp; Entrepreneurship</td>
</tr>
<tr>
<td></td>
<td>Corporate Responsibility</td>
</tr>
<tr>
<td></td>
<td>SME &amp; Family Business</td>
</tr>
<tr>
<td></td>
<td>Human Resources</td>
</tr>
</tbody>
</table>

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 service1. 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 $\kappa$ (Cohen, 1960). The sample selected for evaluation by both annotators is 1200 examples, randomly sampled across languages and topics. The combined $\kappa$ of 0.94 suggests a near-prefect agreement. Table 5 depicts the topic-level $\kappa$.

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

---

1Google 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.

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 (fasttext\(_{LSTM}\)) and a bag-of-embeddings (fasttext\(_{bag}\)) 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

\(^2\)The number of trainable parameters for each model is listed in Table 8 in the Appendix.

\(^3\)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.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALL</td>
<td>NO ENGLISH</td>
</tr>
<tr>
<td>FASTTEXT_BAG</td>
<td>ALL</td>
<td>74.2 ±0.2</td>
</tr>
<tr>
<td>FASTTEXT_BAG</td>
<td>ENGLISH</td>
<td>41.8 ±1.5</td>
</tr>
<tr>
<td>FASTTEXT_BAG</td>
<td>HIGH RESOURCE</td>
<td>70.3 ±11.0</td>
</tr>
<tr>
<td>FASTTEXT_LSTM</td>
<td>ALL</td>
<td>85.4 ±0.4</td>
</tr>
<tr>
<td>FASTTEXT_LSTM</td>
<td>ENGLISH</td>
<td>51.6 ±0.5</td>
</tr>
<tr>
<td>FASTTEXT_LSTM</td>
<td>HIGH RESOURCE</td>
<td>82.4 ±0.6</td>
</tr>
<tr>
<td>sBERT</td>
<td>ALL</td>
<td>73.5 ±0.2</td>
</tr>
<tr>
<td>sBERT</td>
<td>ENGLISH</td>
<td>50.8 ±0.5</td>
</tr>
<tr>
<td>sBERT</td>
<td>HIGH RESOURCE</td>
<td>69.9 ±0.3</td>
</tr>
<tr>
<td>mBERT</td>
<td>ALL</td>
<td>88.6 ±0.3</td>
</tr>
<tr>
<td>mBERT</td>
<td>ENGLISH</td>
<td>58.3 ±0.7</td>
</tr>
<tr>
<td>mBERT</td>
<td>HIGH RESOURCE</td>
<td>84.1 ±0.4</td>
</tr>
<tr>
<td>XLM-R</td>
<td>ALL</td>
<td>90.8 ±0.4</td>
</tr>
<tr>
<td>XLM-R</td>
<td>ENGLISH</td>
<td>68.0 ±1.3</td>
</tr>
<tr>
<td>XLM-R</td>
<td>HIGH RESOURCE</td>
<td>88.6 ±0.4</td>
</tr>
<tr>
<td>MT5</td>
<td>ALL</td>
<td>81.3 ±0.3</td>
</tr>
<tr>
<td>MT5</td>
<td>ENGLISH</td>
<td>50.7 ±1.0</td>
</tr>
<tr>
<td>MT5</td>
<td>HIGH RESOURCE</td>
<td>78.5 ±0.3</td>
</tr>
</tbody>
</table>

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.
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 fastTextLSTM 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, fastTextLSTM 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 $\leq 0.05$) 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, ipn, 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 fastTextLSTM, 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), fastTextLSTM 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-
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.

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., NO ENGLISH, and LOW RES.](image)

The monolingual model’s advantage of languagespecificity 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. 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 mDAFT 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.

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


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


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.


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.

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

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 tokenized our data according to how the fasttext training data was tokenized, using Mecab for Japanese, and the tokenizer from the Europarl preprocessing tools (Koehn, 2005) for the other languages.

Embedding alignment We map monolingual fasttext embeddings trained on Wikipedia and Common-crawl 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. We only keep single-word transla-

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

<table>
<thead>
<tr>
<th>No.</th>
<th>Topic</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Technology</td>
<td>1,088</td>
</tr>
<tr>
<td>2</td>
<td>Industry</td>
<td>1,239</td>
</tr>
<tr>
<td>3</td>
<td>Tax &amp; Accounting</td>
<td>3,371</td>
</tr>
<tr>
<td>4</td>
<td>Finance</td>
<td>1,447</td>
</tr>
<tr>
<td>5</td>
<td>Government &amp; Controls</td>
<td>912</td>
</tr>
<tr>
<td>6</td>
<td>Business &amp; Management</td>
<td>1,991</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>10,045</td>
</tr>
</tbody>
</table>

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

### 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](https://huggingface.co/sentence-transformers/all-mpnet-base-v2)
- **mBERT**: [https://huggingface.co/bert-base-multilingual-cased](https://huggingface.co/bert-base-multilingual-cased)
- **XLM-R**: [https://huggingface.co/xlm-roberta-base](https://huggingface.co/xlm-roberta-base)
- **MT5**: [https://huggingface.co/google/mt5-base](https://huggingface.co/google/mt5-base)

Pre-trained fasttext embeddings can be found at:

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Test</th>
<th>ALL</th>
<th>NO ENGLISH</th>
<th>LOW RESOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FASTTEXT BAG</td>
<td>ALL</td>
<td>78.1 ±0.2</td>
<td>76.7 ±0.8</td>
<td>70.5 ±1.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENGLISH</td>
<td>60.0 ±1.0</td>
<td>52.2 ±1.1</td>
<td>47.7 ±1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIGH RESOURCE</td>
<td>73.6 ±2.4</td>
<td>71.4 ±2.1</td>
<td>52.8 ±1.8</td>
<td></td>
</tr>
<tr>
<td>FASTTEXT LSTM</td>
<td>ALL</td>
<td>83.1 ±0.7</td>
<td>81.3 ±0.8</td>
<td>75.9 ±1.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENGLISH</td>
<td>64.1 ±1.5</td>
<td>55.7 ±1.9</td>
<td>51.6 ±1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIGH RESOURCE</td>
<td>80.4 ±0.4</td>
<td>77.6 ±0.5</td>
<td>60.5 ±1.5</td>
<td></td>
</tr>
<tr>
<td>sBERT</td>
<td>ALL</td>
<td>72.4 ±0.8</td>
<td>66.1 ±1.0</td>
<td>55.3 ±1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENGLISH</td>
<td>51.9 ±0.5</td>
<td>38.4 ±0.8</td>
<td>32.3 ±0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIGH RESOURCE</td>
<td>72.1 ±0.6</td>
<td>65.3 ±0.7</td>
<td>33.0 ±1.5</td>
<td></td>
</tr>
<tr>
<td>mBERT</td>
<td>ALL</td>
<td>87.4 ±0.4</td>
<td>85.0 ±0.4</td>
<td>79.1 ±0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENGLISH</td>
<td>60.4 ±2.4</td>
<td>48.4 ±3.2</td>
<td>48.1 ±2.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIGH RESOURCE</td>
<td>82.9 ±0.5</td>
<td>79.0 ±0.7</td>
<td>52.3 ±2.0</td>
<td></td>
</tr>
<tr>
<td>XLM-R</td>
<td>ALL</td>
<td>89.5 ±0.4</td>
<td>87.8 ±0.5</td>
<td>84.0 ±0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENGLISH</td>
<td>74.9 ±2.2</td>
<td>68.5 ±2.7</td>
<td>67.9 ±1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIGH RESOURCE</td>
<td>87.5 ±0.7</td>
<td>85.3 ±0.8</td>
<td>74.7 ±1.0</td>
<td></td>
</tr>
<tr>
<td>MT5</td>
<td>ALL</td>
<td>83.6 ±0.4</td>
<td>79.7 ±0.5</td>
<td>61.3 ±1.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ENGLISH</td>
<td>56.6 ±0.7</td>
<td>42.9 ±0.8</td>
<td>41.5 ±1.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HIGH RESOURCE</td>
<td>81.1 ±0.6</td>
<td>76.2 ±0.1</td>
<td>43.9 ±0.1</td>
<td></td>
</tr>
</tbody>
</table>

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.