

MFinBERT: Multilingual Pretrained Language Model For Financial Domain

Duong Nguyen*, Nam Cao*, Son Nguyen *[†], Son Ta, Cuong Dinh.
VPS Securities JSC

[†] Corresponding author

anhduongng.1001@gmail.com, chnhust1@gmail.com, son.nv.2421998@gmail.com,
congson1293@gmail.com, dmcksclick5@gmail.com.

Abstract—There has been an increasing demand for good semantic representations of text in the financial sector when solving natural language processing tasks in Fintech. Previous work has shown that widely used modern language models trained in the general domain often perform poorly in this particular domain. There have been attempts to overcome this limitation by introducing domain-specific language models learned from financial text. However, these approaches suffer from the lack of in-domain data, which is further exacerbated for languages other than English. These problems motivate us to develop a simple and efficient pipeline to extract large amounts of financial text from large-scale multilingual corpora such as OSCAR and C4. We conduct extensive experiments with various downstream tasks in three different languages to demonstrate the effectiveness of our approach across a wide range of standard benchmarks.

Index Terms—MFinBERT, Multilingual Language Model, NLP, Fintech, Domain Adaptation, BERT.

I. INTRODUCTION

In recent years, natural language processing (NLP) has been utilized by increasingly more financial technology firms. For instance, having an automated sentiment analysis system that provides market sentiment trends based on daily news would be beneficial as stock prices are heavily influenced by news that affects investors' expectations [12]. In addition, with the growing demand for automated customer service in the financial applications, there has been a growing need for the adoption of news sentiment analysis and question answering systems on chatbots. Consequently, it becomes essential to build language models specific to the financial domain.

The use of transformer-based pre-trained language models has become the standard in NLP these days as they have yielded significant performance gains over previous approaches. Unfortunately, most of these publicly available pre-trained language models are trained on the general domain, which are not well suited for downstream tasks in a variety of specific domains. The common wisdom is that domain adaptation lead to better exploiting in-domain semantic knowledge and ultimately to superior performance on downstream tasks [16], [10]. Specifically for the financial tasks, [1] further pre-train BERT on a subset of news article published by Reuters that contained finance-related keywords. [20], on the other hand, gather several financial communication corpora for both training from scratch and fine tuning BERT. These approaches

have limitations in terms of training data, such as modest sizes and lack of diversity. The problem is further exacerbated when it comes to languages other than English, where such datasets are rarely available. In this paper, we address these challenges by building an efficient pipeline to extract large amounts of financial text from large-scale multilingual corpora such as OSCAR and C4. We name our pretrained language model as Multilingual Financial BERT (MFinBERT) and summarize the contributions of our work as follows:

- We introduce an efficient pipeline that can extract documents containing financial keywords from many publicly available multilingual corpora.
- We publish multilingual pretrained language models in financial domain¹.
- We conduct extensive experiments on sentiment analysis and question answering tasks in three different languages (English, Vietnamese and Lithuania) and analyze the results to demonstrate the effectiveness of our approach.

II. PROPOSED METHOD

In this section, to build our model, we introduce the multilingual data collection and processing flow and the set of keywords that we use to collect raw data. Then we present the architecture used to finetune our language model.

A. Data Filtering from large corpus

A large amount of data is crucial in building a language model. However, financial data are usually private due to customer policy. We can use financial articles as an alternative source, but they are rare and difficult to filter out of the common domain. To address this problem, we introduce a pipeline to filter data from widely available corpus sources (OSCAR, C4). Our filtering is built on top of the data pre-processing of the Huggingface library ². The primary process contains four-part: Filtering by keywords, Deduplicate, Modification of documents, and filtering of records. The overall are shown in Fig.1.

¹<https://huggingface.co/sonnv/MFinBERT>

²https://github.com/bigscience-workshop/data_tooling/tree/master/ac_dc

* These authors contributed equally to this work.

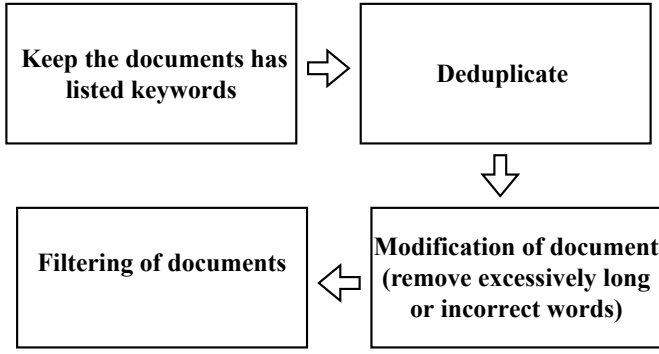


Figure 1: The overall pipeline of data pre-processing and collecting

a) *Modification of the document:* Initially, we built keywords set by the financial experts to remove most of the noisy and irrelevant data. The keywords set is detailed in Table I. Each word in this list satisfies the most specific financial criteria and rarely appears in other fields. After that, we remove duplicate documents using Simhash methods with short documents and suffix array substring deduplication [8] to remove duplicate elements in long documents. The remaining documents will be proceeded to clean with several steps, specifically as follows:

- **Whitespace standarization:** Normalize sentences containing multiple spaces, and carriages return to standard form.
- **Remove long words:** Split the document into words according to the newline character "\n", then for each element split according to "\t", then each sub-element will be split according to the whitespace " ". After that, we strip each word according to the special character and then remove the word with a length more than the threshold. The retained words will be joined back to a normalized form.
- **Remove word with incorrect substrings:** Remove a word if it contains any substring that we consider incorrect. The substrings can be ["http", "www", ".com", "href", "///"]. The goal is to remove links and words related to the page's source code.

b) *Filtering of documents:* The document after modification, will be processed with the steps as follows:

- **Filtering on the character and word repetition ratio:** First, define the length of the repetitions on characters as n . For a document, take the list of $n - grams$ character $L = L_1, L_2, \dots, L_{n-grams}$ and count their frequencies. After that, calculate the total frequencies of the $\sqrt{n-grams}$ elements that have the enormous frequencies, denoting it is $no.Repeat$. If only if $no.Repeat \geq threshold$; remove these documents out of the dataset. We do the same process with word-level.
- **Filtering special characters, stop words ratio:** We pre-defined a list of the special characters and the stop word list for each language. The special characters can be emojis, and punctuation,. It is removed if a document

has a special characters ratio and stop words ratio more significant than a threshold.

- **Filtering by bad words:** In more detail, here is the methodology used to build a list of bad words for a language. Make a list of the most frequent words that appear in pornographic materials for this language. We remove the document with a lousy word ratio more significant than the threshold. The purpose of this filter is not to remove erotic texts but to remove the numerous documents consisting of an accumulation of buzzwords centered on porn, most often without cohesion within the same sentence, which would harm the learning of the model.
- **Filtering on language identification prediction score:** We use the fastText model to detect if the document is written in English. With another language, we do not use these steps. This filter removes documents in which the spoken language changes several times and records of another language, which cannot be correctly analyzed by filters that have been specified with parameters related to a particular language.

Finally, we receive a large well-training dataset after the collecting and pre-processing process from raw data (OSCAR, C4).

B. Multilingual BERT for financial domain: MFinBERT

a) *Architecture:* Our MFinBERT model using the same architecture of BERT-base-multilingual-cased³. Our MFinBERT fine-tuning approach is based on a BERT-based model for fair comparison of number parameters with the previous approach. We finetune from the bert-base-multilingual-cased model to take advantage of previously trained weights and reduce costs.

b) *Pre-training data:* After filtering data from a large corpus, we received 18 GB of English data from uncompressed texts, 15 GB of Vietnamese data from uncompressed texts, and 300 MB of data from Lithuania. The total samples we use are 5.5 billion sub-words with 3.6 billion word tokens.

c) *Otimization:*

- Maximum length: 512 word-piece tokens
- Optimize model : Adam [7]
- Batch size: 128, with 8 Core TPUs v3-8
- Learning rate: $5e - 5$, 5 epochs
- 12-layer, 768-hidden, 12-heads, 110M parameters

III. EXPERIMENTS

The following section describes all the tasks related to the study, and the datasets to be evaluated. Descriptive statistics for the latter are provided in Table II.

A. Downstream task

1) *Text classification:* Text classification is one of the most important task in natural language processing. It is a machine learning technique that assigns a set of predefined categories

³<https://github.com/google-research/bert/blob/master/multilingual.md>

Language	Keywords
English	net profit, dividend stock, gross margin, price target, buy rating, hold rating, sell rating, bull market, bear market, sideways market, sell in mat, fiscall policy, monetary policy, hyperinflation, stagflation, futures contract, balance of trade, trade deficiic, trade surplus
Vietnamese	cổ phiếu, chứng khoán, cán cân thương mại, thâm hụt thương mại, tài khóa, xuất siêu, nhập siêu, grdp, gdp, nợ xấu, cơ cấu nợ, trái phiếu, công nghệ tài chính
Lithuania	Grynasis pelnas, infiacija, vertybiniai popieriai, obligacijos

Table I: The keywords for filtering data

Dataset	Train	Dev	Test	Classes
Financial Phrase Bank	4846	-	-	3
Lithuania Financial News Sentiment	10375	-	-	3
Vietnamese Financial Sentiment	1995	-	-	3
Causality Detection	13,478	-	8,580	2
FiQA 2018-task 2	14166	1238	1706	-

Table II: Descriptive statistics for all the experimental datasets: train and test splits, classes

to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web. To demonstrate the effect of our MFinBERT model on the text classification task, we experimented with two subtasks: Sentiment Analysis and Causality Detection.

Sentiment Analysis is one of the most often used NLP tasks. We chose three distinct datasets in three different languages (English, Lithuania and Vietnam) to compare our models in the finance sector. The first one is Financial Phrase Bank dataset [13] - an English dataset for sentiment classification that consists of 4,840 sentences from financial news that have been annotated for Positive, Negative, and Neutral sentiment by 16 different annotators with financial domain knowledge. Based on the previous research [1], we do 10-fold cross validation for evaluation of the model for this dataset.

The second dataset is Lithuanian Financial News Sentiments dataset [18]. The data were gathered from September 2020 to March 2021, using the “business news” category from the four most popular Lithuanian language news websites. The collected records were highly related to the financial area, where the texts were about financial operations, business projects, investments, political decisions that influenced companies, business expansion, law enforcement, etc. The data in the dataset were labeled by a company; the main field was accounting and business management software development with employees having more than 30 years of experience. The dataset was manually given sentiment (positive, negative, or neutral) by five financial experts. The dataset contained 10375 texts, where 5780 texts were assigned to the class positive (POS), neutral (NEU)—1997 texts, and 2598 texts were negative (NEG). we do 5-fold cross validation for evaluation of the model for this dataset.

Finally, We introduced Vietnamese Financial Sentiment dataset, which was constructed by our team. Firstly, we crawled massive amount of newspapers (documents) from news websites. Then, we defined some key words, which are relevant to

the finance domain in order for data filtering. We only kept documents which composed at least one defined key words. Because the length of a document is large and it could consist of various semantic meanings, we separated it into several paragraph. Finally, four workers carried out labeling data by assigning sentiment for each paragraph (positive, negative and neutral). we experimented 5-fold cross validation for evaluation of the model for this dataset.

For Causality Detection, we used the dataset from the FinCausal shared task 2020 [14] for Document Causality Detection. The dataset is made up of texts collected from Qwan’s 2019 corpus of financial news, with each instance tagged with binary labels indicating whether or not it described a causal relationship. The italics part of (1), for example, was annotated as the source of the GDP decline.

(1) *Things got worse when the Wall came down.* GDP fell 20% between 1988 and 1993.

We are referring to the dataset for subtask 1, which is basically considered as a binary classification task (1 if the text contains a causal link, else 0).

To fine-tune on text classification task, we followed strategy introduced in [2]. We used the last hidden state of the [CLS] token as the aggregate representation and add a fully connected layer as size equal to the number of classes in order for classification purpose.

We used batch size of 8 and fine-tune for 20 epochs for Causality Detection dataset and 50 epochs for all dataset belonging to Sentiment analysis subtask. For each task, we follow the discriminative fine-tuning strategy introduced in [5] for learning rate selection. Particularly, We apply Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and setting learning rate = $5e - 6$ for transformers layers and $1e - 3$ for classifier layer.

2) *Question Answering*: Natural language text question answering (QA) is a fundamental challenge in natural language processing (NLP), with the goal of automatically providing responses to queries about a given short text or passage. Financial QA dataset used in this paper is FiQA Task 2 [11].

FiQA is a dataset that was created for WWW18 conference financial opinion mining and question answering challenge¹². Here we use the data of Task 2, “Opinion-based QA over financial data”. This data included a list of answers and a list of questions (queries) relevant to financial domain. Given a query, our mission is to find the most reasonable answer from answer list for it. This task is considered as information retrieval problem.

Information retrieval is the process of searching and returning relevant documents for a query from a collection. Traditionally, lexical approaches like TF-IDF and BM25 [17] have dominated textual information retrieval. However, this method is time-consuming and can not be able to capture semantic meaning of text for searching process. Recently, there is a strong interest in using neural networks to improve or replace these lexical approaches. This method are capable of capturing semantic matches and try to surpass the (potential) lexical gap. It is be able to map queries and documents in a shared, dense vector space. This allowed the document representation to be pre-computed and indexed. A bi-encoder neural architecture based on pre-trained Transformers has shown strong performance for various open-domain question-answering tasks [3], [6], [6], [9].

In our paper, we used **DPR** [6] for QA task. This model is a two-tower bi-encoder trained with a single BM25 hard negative. The two encoders used the same language model to produce contextualized presentations of both question and answer, which will be parallelly updated during training process. Each query was given true answer(s) or positive samples and negative samples are generated as follow: firstly, applying BM25 to find top k (hyper-parameters, in this paper we set $k = 10$) document from answer collections which have the highest similarity score with the positive samples. Documents which is not the ‘real’ answers will be setted as negative samples. During training process, model aim to maximize the similarity score (cosine) between the representation of query and its positive sample as well as minimize score between query and negative sample. Model is trained for 3 epoches, and loss function we use is Multiple Negative Ranking Loss [4]. Code for this model can be found at: <https://github.com/beir-cellar/beir>

B. Baseline and comparison models

We fine tune our based language model: bert-based-multilingual-cased in all datasets to compare with our MFinBERT model. We also compared the performance of our model with the result reported in previous research, particularly [1] and [20] for Financial Phrase Bank dataset, [15] for Causality Detection dataset, [18] for Lithuanian Financial News Sentiment dataset and [19] for question answering. For the Vietnamese Financial Sentiment dataset, we only fine tune bert-base-multilingual-cased as base model.

C. Evaluation Metrics

Our experiments consisted of two kinds of tasks: classification and question answering. For evaluation of classification

models, in order to compare with previous works, we used accuracy and macro F1 average in Financial Phrase Bank, Lithuanian Financial News Sentiment and Vietnamese Financial News Sentiment dataset. In Causality Detection dataset, we reported Macro F1 average and Micro F1 average score. For evaluation question answering models, we used two most popular metrics applying for evaluating information retrieval: Mean Reciprocal Rate (MRR@ k) and Normalised Cumulative Discount Gain (nDCG@ k).

IV. RESULTS AND DISCUSSION

The full results about text classification task are shown in table III. We observed that our model outperformed other baseline method in each dataset. There is a significant improvement in Vietnamese dataset, increasing by approximately 6% in accuracy and 5% in macro-F1. On English and Lithuanian dataset, MFinBERT performed similarly to base model, only showing small improvements.

Table IV and V show our results on financial QA task. From experiments, we can obviously witness that MFinBERT are much better than the previous models. Our model significantly outperformed all baseline ones, and achieved approximately 5% in all metrics in nDCG and 6% for MRR. The experimental results are highly effective and encouraging.

On all the financial datasets, MFinBERT achieved better results, which proves the effectiveness of our proposed language model. Experimental results highlight the importance of the financial domain-specific corpora pre-trained design. This is proved most clearly on Vietnamese dataset. Vietnamese language is our nation one, so we understand deeply about semantic meaning and to be easy to define valuable financial keywords for data filtering. Fintuning in a high-quality dataset of the financial domain made our model capture knowledge of this field hence it boost performance of model in Vietnamese downstream task considerably. For Lithuanian dataset, because of our restriction about language understanding, we only determined few Lithuanian financial keywords. As a result, our Lithuanian financial corpus for retraining language model is not very large (100 MB), therefore the performance of fine-tuning model is slightly increase in comparison to base model.

We measured the effect of data size for fine-tune language model on the performance of the classifier. We compare three model: 1) FinBERT [1], 2) FinBERT [20] and our MFinBERT. In order to pre-train FinBERT [1], authors use a financial corpus called TRC2-financial. It is a subset of Reuters’TRC24, which consists of 1.8M news articles that were published by Reuters between 2008 and 2010. Data was filtered for some financial keywords in order to make corpus more relevant. The resulting corpus, TRC2-financial, includes 46,143 documents with more than 29 million words and nearly 400K sentences. In [20], corpora was collected from financial website. The total size of of this corpora is approximately 4.9 billion tokens. Meanwhile, our dataset contains 3.6 billion words in total corresponding to 5.5 billion tokens. As compare to these dataset, the size of both datasets is smaller than our English corpus, as a result, our model get better performance.

Dataset	Model	Accuracy	Macro-F1	Micro-F1
Financial Phrase Bank	FinBERT [1]	0.86	0.84	-
	bert-base-cased [20]	0.76	-	-
	FinBERT [20]	0.86	-	-
	bert-base-multilingual-cased	0.86	0.84	-
	MFinBERT (ours)	0.87	0.85	-
Lithuanian Financial News Sentiment	Naive Bayes [18]	0.71	-	-
	bert-base-multilingual-cased	0.69	0.59	-
	MFinBERT (ours)	0.72	0.62	-
Vietnamese Financial Sentiment	bert-base-multilingual-cased	0.76	0.73	-
	MFinBERT (ours)	0.81	0.79	-
Causality Detection	FinBERT [15]	-	0.95	0.80
	bert-base-multilingual-cased	-	0.96	0.80
	MFinBERT (ours)	-	0.96	0.81

Table III: Experimental Results on text classification task

Model	NDCG@1	NDCG@3	NDCG@5	NDCG@10	NDCG@100	NDCG@1000
bert-base-uncased [19]				0.112		
FinBERT [1]	0.131	0.129	0.137	0.157	0.226	0.268
Bert-based-multilingual-cased	0.133	0.120	0.127	0.146	0.203	0.247
MFinBERT (ours)	0.176	0.165	0.170	0.192	0.259	0.303

Table IV: Experimental results on the test set for the FiQA question answering dataset with NDCG score.

Model	MRR@1	MRR@3	MRR@5	MRR@10	MRR@100	MRR@1000
FinBERT [1]	0.131	0.175	0.186	0.198	0.210	0.211
Bert-based-multilingual-cased	0.133	0.167	0.177	0.187	0.197	0.198
MFinBERT (ours)	0.176	0.223	0.232	0.245	0.256	0.257

Table V: Experimental results on the test set for the FiQA question answering dataset with MRR score.

V. CONCLUSION

This paper has implemented a multilingual pre-trained model for a financial domain by fine-tuning the bert-based-multilingual-cased model. We introduce a pipeline to collect and clean data for the domain adaptation model. The pipeline is easy to implement and costs less than the crawl data approach. Then, we fine-tune this model and evaluate various downstream tasks (news sentiments, QA) in three languages: English, Vietnamese, and Lithuania and compare it with previous approaches. Our results show that the proposed model performs significantly better solution quality with the same amount of parameters. We will experiment with more language and more downstream tasks for future work. After that, we will use a more efficient multilingual model for fine-tuning such as XLM-Roberta.

ACKNOWLEDGMENT

This research is supported by VPS Securities JSC. For the hardware to training model, we used TPU v3-8 thank to the support of TPU Research Cloud.

REFERENCES

- [1] Dogu Araci. “Finbert: Financial sentiment analysis with pre-trained language models”. In: *arXiv preprint arXiv:1908.10063* (2019).
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [3] Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. “MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models”. In: *arXiv preprint arXiv:2005.02507* (2020).
- [4] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. “Efficient natural language response suggestion for smart reply”. In: *arXiv preprint arXiv:1705.00652* (2017).
- [5] Jeremy Howard and Sebastian Ruder. “Universal language model fine-tuning for text classification”. In: *arXiv preprint arXiv:1801.06146* (2018).
- [6] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. “Dense passage retrieval for open-domain question answering”. In: *arXiv preprint arXiv:2004.04906* (2020).
- [7] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [8] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. “Deduplicating training data makes language models better”. In: *arXiv preprint arXiv:2107.06499* (2021).
- [9] Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. “Embedding-based zero-shot retrieval through query generation”. In: *arXiv preprint arXiv:2009.10270* (2020).

- [10] Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sad-eque, Guergana Savova, and Timothy A Miller. “Does BERT need domain adaptation for clinical negation detection?” In: *Journal of the American Medical Informatics Association* 27.4 (2020), pp. 584–591.
- [11] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. “Www’18 open challenge: financial opinion mining and question answering”. In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 1941–1942.
- [12] Burton G Malkiel. “The efficient market hypothesis and its critics”. In: *Journal of economic perspectives* 17.1 (2003), pp. 59–82.
- [13] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. “Good debt or bad debt: Detecting semantic orientations in economic texts”. In: *Journal of the Association for Information Science and Technology* 65.4 (2014), pp. 782–796.
- [14] Dominique Mariko, Estelle Labidurie, Yagmur Ozturk, Hanna Abi Akl, and Hugues de Mazancourt. “Data Processing and Annotation Schemes for FinCausal Shared Task”. In: *arXiv preprint arXiv:2012.02498* (2020).
- [15] Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Churen Huang. “Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks”. In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. 2021, pp. 37–44.
- [16] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. “Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification”. In: *arXiv preprint arXiv:1908.11860* (2019).
- [17] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.
- [18] Rokas Štrimaitis, Pavel Stefanovič, Simona Ramanauskaitė, and Asta Slotkienė. “Financial context news sentiment analysis for the Lithuanian language”. In: *Applied Sciences* 11.10 (2021), p. 4443.
- [19] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. “BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models”. In: *arXiv preprint arXiv:2104.08663* (2021).
- [20] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. “Finbert: A pretrained language model for financial communications”. In: *arXiv preprint arXiv:2006.08097* (2020).