A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain

Yusuf Arslan
yusuf.arslan@uni.lu
University of Luxembourg
Luxembourg

Kevin Allix
University of Luxembourg
Luxembourg
kevin.allix@uni.lu

Lisa Veiber
University of Luxembourg
Luxembourg
lisa.veiber@uni.lu

Cedric Lothritz
University of Luxembourg
Luxembourg
cedric.lothritz@uni.lu

Tegawendé F. Bissyandé
University of Luxembourg
Luxembourg
tegawende.bissyande@uni.lu

Jacques Klein
University of Luxembourg
Luxembourg
jacques.klein@uni.lu

Anne Goujon
BGL BNP Paribas
Luxembourg
anne.goujon@bgl.lu

ABSTRACT

Neural networks for language modeling have been proven effective on several sub-tasks of natural language processing. Training deep language models, however, is time-consuming and computationally intensive. Pre-trained language models such as BERT are thus appealing since (1) they yielded state-of-the-art performance, and (2) they offload practitioners from the burden of preparing the adequate resources (time, hardware, and data) to train models. Nevertheless, because pre-trained models are generic, they may underperform on specific domains. In this study, we investigate the case of multi-class text classification, a task that is relatively less studied in the literature evaluating pre-trained language models. Our work is further placed under the industrial settings of the financial domain. We thus leverage generic benchmark datasets from the literature and two proprietary datasets from our partners in the financial technological industry. After highlighting a challenge for generic pre-trained models (BERT, DistilBERT, RoBERTa, XLNet, XLM) to classify a portion of the financial document dataset, we investigate the intuition that a specialized pre-trained model for financial documents, such as FinBERT, should be leveraged. Nevertheless, our experiments show that the FinBERT model, even with an adapted vocabulary, does not lead to improvements compared to the generic BERT models.

CCS CONCEPTS

• **Applied computing** → Text processing.

KEYWORDS

BERT, FinBERT, financial text classification

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451375

ACM Reference Format:
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 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3442442.3451375

1 INTRODUCTION

Financial institutions receive, process, generate, and send a considerable amount of financial documents on a daily basis, including but not limited to policy descriptions, financial statements, prospectuses, customer reports. The first step towards the exploitation of these large volumes of documents is to ensure their proper classification. In most cases, however, as observed in several firms' workflows, document classification remains a significantly manual process. Yet, manual classification is known to be error-prone, with catastrophic implications in the application in industrial settings: at JP Morgan, contract interpretation errors stem from manual processing errors and lead to 80% of the firm's loan servicing errors [51]. Automated text classification has been extensively investigated in the literature [3]. Although it remains, so far, scarcely adopted by financial institutions [25], potentially due to the substantial effort that must be undertaken to train and keep machine learning models up-to-date and accurate. Nevertheless, in recent years, Pre-trained Language Models (PLMs) have offered the research and practice communities with a breakthrough for adopting automated Natural Language Processing (NLP) techniques, even with limited time and computational resources. PLMs are further appealing since they have yielded state-of-the-art performance[1] in many NLP subtasks [43]. However, their performance for the task of full-document classification is not yet perceived as reliable as for other sub-tasks, making our project partners in the finance industry reluctant to deploy them in their highly regulated environments.

[1] https://gluebenchmark.com/leaderboard

Online repositories share a large variety of released PLMs. For instance, we can find more than 5 000 models under community models[2]. These PLMs are generated under various settings, although with limited information to their applicability to industrial datasets and tasks. Therefore, we argue for the present need to validate existing models for different tasks and domains rather than incessantly spending resources to build and release model variants.

Pre-trained models increasingly appear as the new standard of industry best practices. Indeed, leveraging pre-trained models in the industry offers several benefits, including (1) the ease of integration within a production workflow, (2) the possibility to achieve high performance quickly, i.e., without too many fine-tuning iterations, (3) the opportunity to escape the need of huge labeled datasets for training, as well as (4) their applicability across various use cases and tasks. Furthermore, in practice, training deep neural networks from scratch generally requires specialized computing architecture (e.g., Tensor Processing Units (TPU)) and incur significant costs. Indeed, training the BERT-based model has been evaluated in the literature to cost approximately $7 000 for 4 days usage of 4 Clouds TPUs [40]. Moreover, the carbon footprint generated by the redundant process remains problematic, as is the use of public cloud resources to process critical business, which can raise data confidentiality and privacy concerns for companies.

Nevertheless, recent works begin to formulate some criticisms over the performance of PLMs on specific datasets [28, 31]. A summary insight of these experiments is that PLMs should be used with great care as their performance may be task-dependent and dataset-dependent. Therefore, it is important to investigate the capabilities of existing models and architectures in these dimensions so as to identify improvement directions.

In the scope of our collaborative projects with partners from the financial technology industry, we undertook the task of assessing pre-trained language models for several NLP tasks. In this paper, we discuss our experiments and the yielded insights for the case of document classification. We notably assess the value of leveraging a specialized model against the use of generic models pre-trained on common text data. To the best of our knowledge, this work is the first to study PLMs for multi-class text classification for the financial domain. Concretely, we address the following Research Questions (RQs):

RQ1: What is the performance of generic pre-trained language models on the task of multi-class text classification? We perform the experiments on datasets from both financial and non-financial domains to draw a comparison baseline.

RQ2: Does FinBERT (a financial domain-specific language model) outperform the generic pre-trained language models on datasets from financial domains for the task of multi-class text classification?

RQ3: To what extent does the vocabulary impact the performance of pre-trained language models in the multi-class text classification task? We investigate the overlap between the vocabularies and explore the performance that can be gained with a custom vocabulary.

This paper is organized as follows. Section 2 overviews some related work while Section 3 provides some background information about language models. We present our experimental setup in Section 4 and discuss the results in Section 5 as well as an enumeration of some threats to validity. Finally, Section 6 concludes the paper.

[2]https://huggingface.co/models

2 RELATED WORK

In the last decade, many proposed NLP approaches exploited neural networks, which use task-specific data and word embeddings such as word2vec [29] and GloVe [33]. Transformer models, which remove recurrence as well as convolution and depend on attention, were introduced by Vaswani et al. [42]. It led to a paradigm shift in NLP domain such that pre-trained deep language representation models come to play as a commonly-used type of natural language models.

2.1 Pre-trained Language Models (PLMs)

PLMs are language models that have been trained with a large dataset while remaining agnostic to the specific tasks they will be employed on. In practice, to leverage PLMs, the last output layers must be adapted to the task: this is referred to in the literature as the *fine-tuning* step.

OpenAI GPT [34], BERT [10], XLNet [47] and XLM [8] are examples of pre-trained models that can be fine-tuned to various NLP tasks. PLMs received huge attention after BERT achieved state-of-the-art results on 11 NLP tasks [10]. Variants of the BERT model and other PLMs can be found in online repositories. Currently, more than 5 000 models have already been made available to the community. These models can be broadly categorized as either: a) adaptation to a specific task and/or a specific domain, or b) optimization, where the goal is to improve the core of the model or reduce its computational cost.

While BERT achieves excellent performance in several NLP subtasks, several researchers focused on creating PLMs specifically adapted to the context of a given specific domain, usually by either fine-tuning or fully re-training BERT—*pre-training* in BERT terminology—on another corpus. Accordingly, approaches have been proposed for biomedical language [20], scientific papers [7], clinical notes [4, 14] and financial news [6].

Concurrently, others have experimented to adapt PLMs to tasks not originally evaluated by BERT authors. Such processes usually only involve fine-tuning BERT, being much less computationally expensive than pre-training BERT on another corpus. Adhikari et al. [2] propose to fine-tune BERT to yield a model able to classify a full document, while Lee and Hsiang [19] tackle the problem of classifying patents.

Other research has focused on optimizing PLMs, known to be expensive, and generally hard to pre-train. For instance, distilBERT [38] proposes a modified trade-off between pre-training and fine-tuning, which allows to obtain a smaller model, easier to train, while conserving most of the performance of the original BERT model. On the other hand, Liu et al. [24] argue that BERT requires *more* training, and proposed RoBERTa, a model trained for longer than the original BERT. Their study shows that RoBERTa generally outperforms BERT.

2.2 Evaluation and Limitations of PLMs

PLMs being relatively recent, an active field of research is devoted to uncovering and documenting their limitations. Niven and Kao [31] thoroughly examine BERT accuracy on the Argument Reasoning Comprehension Task. They show that the results of BERT on this task can be accounted for by taking advantage of spurious statistical cues in the dataset. This paper claims that all models achieve random accuracy on adversarial datasets, and suggests the use of adversarial datasets as a standard in the future.

McCoy et al. [28] investigate *why* machine learning systems (including BERT) perform well on a given test set. They found that BERT (and other models) performance may not generalize on other corpora. Schick and Schütze [39] show that language models may struggle to deal with rare words despite being trained on a big amount of data. Furthermore, they showed that the frequency of words is highly important in language models understanding. It can be concluded that datasets with a high number of unique words can be quite challenging for language models.

Sun et al. [41] make detailed experiments on BERT and suggest several techniques to improve the results on text classification task. Yu et al. [50] propose a BERT-based model for text classification to utilize more task-specific knowledge and achieve better results on multi-classification task. Yeung [49] inserts legal domain vocabulary to BERT, reports no improvement and explains their findings by the high overlap between vocabularies. Elwany et al. [11] investigate BERT on large legal corpora and report improvement after fine-tuning on the legal domain.

Li et al. [22] investigate BERT-based models for entity normalization task for biomedical and clinical domain. The study does not detect statistical significance between biomedical and clinical domain, and concludes that the domain effect on models is not statistically significant if the domains of the models are close, while domain effect becomes more visible on distant domains. Peng et al. [32] conduct an empirical study on BERT and its variations on biomedical and clinical domain, and show that fine-tuning models outperform state-of-the-art transformer models.

Like this related work, our paper aims to contribute to this growing literature of empirical evaluations of existing PLMs.

2.3 Text classification

Text classification is one of the classical tasks in NLP. Numerous methods have been proposed to tackle this task, including but not limited to, the use of Naïve Bayes [12, 16, 27, 36, 52], support vector machines [35], random forest [46], hierarchical attention networks [48] and convolutional neural networks [15, 18]. Text classification task can have four levels of granularity, based on text size, which are document level, paragraph level, sentence level and sub-sentence level [17]. Another important aspect is the type of classification, as classification can be either binary (i.e., either a text is member of a group or not), multi-class (i.e., only one among several possible classes), or multi-label (i.e., each input can be associated with several classes).

Multi-class text classification, which is the focus task of our work, has been investigated by several research works in the literature. Li and Vogel [21] improve multi-class classification by using subclass information and present their results on the *20News* dataset, which is one of the datasets used in this paper. Damaschk et al. [9] inspect multi-class classification on datasets that contain unbalanced classes with noisy examples. They conclude that further pre-processing of data, such as removing noisy examples and settling the unbalanced classes, improves the results. Anne et al. [5] classify patent documents to multi-classes and improve results by removing miss-classified files from the training dataset and injecting synthetic data to reduce data imbalance. Lim [23] examines various machine-learning approaches for multi-class text classification on a specific domain of legal documents; One of the important challenges the study faced is this lack of labeled data, which is one of the problems in domain-specific studies.

Overall, although text classification has been studied with various approaches, the literature is limited in terms of works that investigate the use of PLMs for the specific task of text classification.

3 BACKGROUND

PLMs are complex systems, and they significantly differ from previous approaches in both their inner workings and in their usage. In this section, we introduce several key concepts that are fundamental to understanding experiments with PLMs[3]. We focus on the most prominent PLM, BERT, but all BERT-based approaches use the same concepts.

To obtain a high-performance PLM, Devlin et al. [10] combined various building blocks, all of which may contribute to BERT improvements over previous approaches.

3.1 Representation of Text

Before being fed to any machine learning algorithm, textual data must be brought to a suitable form. BERT passes its text input through three layers to transform each token of the input into a vector representation. First, the input text is tokenized, and special [*CLS*] and [*SEP*] tokens are added at the beginning and end of each input. Then, tokens are passed to an embedding layer, and tokenization using WordPiece [45] is performed to generate a vocabulary that contains all English characters, and the most common words and subwords found in the training corpus. This layer transforms each token into a 768-dimensional vector representation. BERT further processes the text to consider positional information and to make the neural-network compatible with BERT training method (discussed below).

3.2 Pre-Training

BERT builds its Language model through a first phase of training, named *pre-training*. This pre-training is performed on the two tasks of "Masked Language Modeling" (MLM) and "Next Sentence Prediction" (NSP). In the MLM phase, the BERT approach randomly masks (i.e., replaces) some of the words in each input with a special token, and then attempts to predict the original value of the masked words. In the NSP phase, the model is fed pairs of sentences as input. The objective is to predict correctly whether the second sentence in the pair is the following sentence in the original document, or is unrelated. BERT NSP training phase uses 50% of the input pairs from the original document, while the second sentences of remaining pairs are chosen randomly from the original document.

[3] Our descriptions here are vastly over-simplified. We refer interested readers to [10]

The insights of Devlin et al. [10] are that the MLM training would *teach* BERT to model relationships between words, while the NSP would let BERT *learn* relationships between sentences, both training tasks complementing each other to build a task-agnostic language model that is aware of relationships between words and between sentences.

3.3 Fine-Tuning

After the pre-training, BERT is not yet ready to be used for standard NLP tasks. Instead, it must be adapted to the task at hand via a Fine-Tuning phase. In practice, this fine-tuning will train the last layers of the neural network to leverage the language model (captured in the other layers of the neural network) to perform the task. Typically, users of BERT would only need to perform this fine-tuning phase, which requires orders of magnitude less computational power than a full pre-training.

3.4 Vocabulary

PLMs are highly adaptable to different domains and tasks. One of the reasons for their flexibility is that they address the vocabulary problem by using subword tokenization methods. BERT extracts subword tokens in the form of WordPiece [45] tokens. Each input word is split until it matches one of the tokens in BERT's WordPiece vocabulary. BERT vocabulary, which contains 30 522 words and subwords, is constructed by using the frequency of sequences of characters in the BERT corpus [30]. This approach may have drawbacks on niche domains like finance, law, and science because of the high number of words unique to this domain. Several studies on specific domains address this drawback by employing a custom vocabulary, that is constructed on a domain-specific dataset. Nevertheless, there is no consensus on the improvement achieved by using a custom vocabulary between the results of those studies [7, 49].

4 EXPERIMENTAL SETUP

In this section, we describe the PLMs, their parameters, and the datasets we use to investigate our research questions.

4.1 Replication of pre-trained Language Models

In our experiments, we use HuggingFace's Transformers library[4] [44] as the source for all PLMs. As described in Section 3, all PLMs require fine-tuning to be adapted to the task at hand. We hence fine-tuned all PLMs so they could be used for multi-class text classification. In all our experiments, we use the same parameters for fine-tuning: train batch size of 16, evaluation batch size of 16, maximum sequence length of 128, and adam learning rate of $4e^{-5}$. We also perform our experiments for 1, 3, and 5 fine-tuning epochs. Those values are among the values that "work well across all tasks" according to Devlin et al. [10].

[4]https://github.com/huggingface/transformers

4.2 Other approaches evaluated

In addition to the original BERT, our work compares the performance of several other PLMs. Here, we briefly present those approaches.

DistilBERT [38] is presented as a "smaller, faster, cheaper, and lighter" (distilled) version of BERT. It is 60% faster than BERT, reducing the size of the BERT model by 40%, while keeping 97% of its language understanding capability.

RoBERTa [24] is designed by re-evaluating and modifying design decisions of BERT. RoBERTa manages to improve the performance of BERT by pre-training longer than BERT, with a larger batch size, modifying the MLM pre-training, and by skipping the NSP pre-training phase.

XLNet [47] uses a generalized auto-regressive pre-training method rather than the auto-encoder based pre-training of BERT. XLNet outperforms BERT on a set of 20 NLP tasks, including text classification.

XLM [8] is modified BERT tailored to specifically address two tasks, namely, cross-lingual classification and machine translation. XLM uses Byte-Pair Encoding (BPE) instead of words or characters encoding, so as to increase the shared vocabulary between languages. It trains BERT with dual-language input to learn cross-language context. It further initializes pre-trained BERT together with translation of model embeddings to improve Back-Translation.

FinBERT [5][6] is a BERT-based PLM dedicated to the financial domain. It brings additional pre-training to specialize BERT on the financial domain by using a subset of Thomson Reuters Text Research Collection [37]. TRC2-financial contains 46 143 documents and approximately 400K sentences. FinBERT also fine-tunes its model for financial sentiment classification by using Financial PhraseBank [26]. We have included FinBERT in our experiments for multi-class classification since FinBERT is specifically targeted at our domain of interest.

4.3 Datasets

In this study, we perform our experiments on four datasets. The *20News* and *BBC* datasets are already used and available to the NLP research community. The *BBC* dataset comes from BBC News [13]. It comprises 2225 articles, which are labeled under one of the five categories, namely: *business*, *entertainment*, *politics*, *sport*, or *tech*. The training set contains 1490 news articles and the test set contains 735 articles. The *20News* dataset is a collection of 18 846 newsgroup documents with 20 classes available online [1].

For this study, we collected two datasets (named here *Proprietary-1* and *Proprietary-2*). These two datasets are obtained from two different European financial institutions that manage large numbers of debt and fund securities. These proprietary datasets contain both public and confidential text documents related to those securities, and thus cannot be made public. The two proprietary datasets are real-world extracts of the typical inflow of documents that financial institutions need to classify before further processing such as data extraction can proceed. We note that from both an internal business perspective and from a regulatory standpoint, there is a strong emphasis on the correctness of the document processing pipeline, regardless of it being manual or automatic.

[5]FinBERT can be found at https://github.com/ProsusAI/finBERT

Indeed, financial institutions must, by law, act when regulatory documents are emitted for security. Failure to do so can lead to major business impact, coupled with potential fines from their local finance regulation body; Repeated offenses could even lead to the loss of their license to operate on financial markets.

The *Proprietary-1* dataset contains 22 323 financial documents with eleven classes. The document classes and the number of documents in each class can be seen in Table 1.

Document Number	Document Class
72	Annual Financial Statement
1808	Base Prospectus
11 332	Final Terms
462	Listing Particulars
2750	Other
181	Registration Document
599	Securities Note
2606	Series Prospectus
105	Standalone
111	Summary
2297	Supplement

Table 1: Proprietary-1 Dataset

The *Proprietary-2* dataset has six classes and 1135 documents in it. The document classes and the number of documents in each class can be seen in Table 2.

Document Number	Document Class
120	Other Third Party Document
148	Registration Document
259	Collective Commitment
54	Securities Note
290	Basic Program Prospectus
264	Unit Prospectus

Table 2: Proprietary-2 Dataset

We note that while documents from both *Proprietary-1* and *Proprietary-2* serve the same purpose of fully describing the events and life-cycle of securities, they do not have exactly overlapping classes nor the same number of classes. This is explained by the fact that regulatory document types are defined by national finance regulation bodies, and the two financial institutions we obtained our datasets from, operate in different countries. A summary of our datasets is presented in Table 3.

5 RESULTS AND DISCUSSION

In this section, we will first answer the three research questions (RQs) that we formulated in the Introduction (cf. Section 1). Then, we discuss the threats to validity related to our study.

5.1 Answers to the Research Questions

RQ1: What is the performance of generic pre-trained language models on the task of multi-class text classification?

To draw a comparison baseline, we apply the 5 generic PLMs introduced previously (i.e., BERT, DistilBERT, RoBERTa, XLM, and XLNet) on the four datasets presented in Section 4.3. We recall that both *20News* and *BBC* contain non-financial documents, while Proprietary-1 and Proprietary-2 only contain financial documents.

Table 4 presents the precision, recall, and F1-score measures (respectively noted P, R and F1 in the table) obtained when experimentally applying the 5 PLMs using 10-fold cross-validation. In the initial BERT paper [10], the authors have empirically validated the number of epochs which should be used in BERT-based experiments. They have shown that results are not improved after 5 epochs. Following this, the scores presented in Table 4 are detailed for 1, 3 and 5 epochs. On the *BBC* dataset, the best precision, recall and F1 score are achieved by RoBERTa for three epochs. The performance scores are extremely high with 0.99 for each metrics. RoBERTa is performing the best on the four datasets. However, other approaches are very close and often reach the performance levels of RoBERTa. Indeed, on the *20News* dataset, except DistilBERT, the other approaches perform as well as RoBERTa, and on the *Proprietary-1* dataset, all approaches perform identically.

On the two smallest datasets, *BBC* and *Proprietary-2*, the generic PLMs perform extremely well, F1-scores being 0.99 and 0.97 respectively, while on the two biggest datasets, *20News* and *Proprietary-1* (which are an order of magnitude –10 times– bigger than the small ones), the performance drops. In particular, it is noteworthy that on the *Proprietary-1* dataset, the performance of all the approaches is significantly lower than on other datasets. Further investigations reveal that among the 11 classes of documents of the *Proprietary-1* dataset, three specific classes of financial documents lead to low performance, suggesting that they may be difficult for the generic PLMs.

RQ1 Answer: While the performances of the 5 generic PLMs are generally very close when applied to a given dataset, RoBERTa is performing the best.

The performances of the generic PLMs are (very) high, except on the biggest dataset of financial documents where we can see a degradation of the performance (less than 0.90 of F1-score).

RQ2: Does FinBERT outperform the generic pre-trained language models on datasets in financial domains for the task of multi-class text classification?

Following our experiments in RQ1, the results suggested that specific datasets of financial documents may not be properly classified with generic PLMs. We thus proceed to investigate the possibility to use a financial domain-specific language model to improve the classification scores on financial datasets. To that end, we compare FinBERT (a domain-specific language model which is further pre-trained and fine-tuned on financial data) against RoBERTa (which yielded the best results in the previous experiment). Table 5 presents the results for one, three and five epochs. Again the precision, recall, and F1-score are computed using ten-fold cross-validation.

As expected, on non-financial datasets (*BBC* and *20News*), FinBERT is not performing better than RoBERTa. However, in contrast, we would expect FinBERT to outperform RoBERTa on both *Proprietary-1* and *Proprietary-2* financial datasets. As shown in Table 5, that is not the case. Indeed, at best, FinBERT reaches the same level of performance as RoBERTa.

	# classes	# documents	# sentences	# words	# unique words	avg. # words
BBC	5	2225	39 697	472 483	23 435	212
20News	20	18 846	307 953	3 234 347	213 673	171
Proprietary-1	11	22 333	12 825 637	306 833 891	2 897 642	13 745
Proprietary-2	6	1135	1 198 371	30 129 324	578 234	26 545
TRC2-financial	-	46 143	400K	29M	-	-
Financial Phrasebank	3	-	4845	63 883	10 445	13
FiQA	[-1,1]	-	1174	12 122	4459	9

Table 3: Datasets Statistics

Epoch		1			3			5		
Datasets	Model	P	R	F1	P	R	F1	P	R	F1
BBC	BERT	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	DistilBERT	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98
	RoBERTa	0.98	0.98	0.98	**0.99**	**0.99**	**0.99**	0.98	0.98	0.98
	XLM	0.89	0.88	0.88	0.97	0.97	0.97	0.97	0.97	0.97
	XLNet	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98
20News	BERT	0.85	0.85	0.85	0.92	0.92	0.92	**0.93**	**0.93**	**0.93**
	DistilBERT	0.82	0.82	0.82	0.90	0.90	0.90	0.91	0.91	0.91
	RoBERTa	0.84	0.84	0.84	0.92	0.91	0.91	**0.93**	**0.93**	**0.93**
	XLM	0.89	0.89	0.89	0.92	0.92	0.92	**0.93**	**0.93**	**0.93**
	XLNet	0.85	0.85	0.85	0.91	0.91	0.91	**0.93**	**0.93**	**0.93**
Proprietary-1	BERT	0.86	0.88	0.87	**0.88**	**0.89**	**0.88**	**0.88**	**0.89**	**0.88**
	DistilBERT	0.86	0.88	0.87	**0.88**	**0.89**	**0.88**	**0.88**	**0.89**	**0.88**
	RoBERTa	0.83	0.86	0.83	**0.88**	**0.89**	**0.88**	0.87	0.88	0.87
	XLM	0.81	0.84	0.81	**0.88**	**0.89**	**0.88**	0.83	0.86	0.83
	XLNet	0.83	0.86	0.82	**0.88**	**0.89**	**0.88**	**0.88**	**0.89**	**0.88**
Proprietary-2	BERT	0.70	0.81	0.74	0.95	0.95	0.95	0.96	0.96	0.95
	DistilBERT	0.72	0.84	0.78	0.96	0.96	0.96	**0.97**	0.96	0.96
	RoBERTa	0.89	0.89	0.87	0.95	0.95	0.95	**0.97**	**0.97**	**0.97**
	XLM	0.71	0.70	0.70	0.91	0.91	0.91	0.94	0.94	0.93
	XLNet	0.91	0.90	0.89	0.94	0.94	0.93	**0.97**	0.96	0.96

Table 4: Results on both financial and non-financial datasets. In bold are the best results for each dataset.

Epoch		1			3			5		
Datasets	Model	P	R	F1	P	R	F1	P	R	F1
BBC	RoBERTa	0.98	0.98	0.98	**0.99**	**0.99**	**0.99**	0.98	0.98	0.98
	FinBERT	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96
20News	RoBERTa	0.84	0.84	0.84	0.92	0.91	0.91	**0.93**	**0.93**	**0.93**
	FinBERT	0.86	0.86	0.86	0.92	0.92	0.92	**0.93**	**0.93**	**0.93**
Proprietary-1	RoBERTa	0.83	0.86	0.83	**0.88**	**0.89**	**0.88**	0.87	0.88	0.87
	FinBERT	0.86	0.88	0.87	**0.88**	**0.89**	**0.88**	**0.88**	**0.89**	**0.88**
Proprietary-2	RoBERTa	0.89	0.89	0.87	0.95	0.95	0.95	**0.97**	**0.97**	**0.97**
	FinBERT	0.74	0.82	0.76	0.96	0.95	0.95	**0.97**	0.96	0.96

Table 5: Comparison of FinBERT against RoBERTa

One possible explanation to these results is that, even if FinBERT has been pre-trained and fine-tuned with financial text data, the specific documents that were leveraged in FinBERT may still be significantly different from the documents contained in *Proprietary-1* and *Proprietary-2*. We will explore this hypothesis in the next RQ.

RQ2 Answer: On the specific task of multi-classification of financial documents, our experiments show that FinBERT, which

Epoch		1			3			5		
Datasets	Model	P	R	F1	P	R	F1	P	R	F1
Proprietary-1	FinBERT	0.86	0.88	0.87	**0.88**	**0.89**	**0.88**	**0.88**	**0.89**	**0.88**
	FinBERT-Custom	0.85	0.87	0.85	**0.88**	0.88	**0.88**	**0.88**	**0.89**	**0.88**
Proprietary-2	FinBERT	0.74	0.82	0.76	**0.96**	0.95	0.95	**0.97**	**0.96**	**0.96**
	FinBERT-Custom	0.70	0.79	0.74	0.94	0.94	0.94	0.95	0.95	0.95

Table 6: Comparison of FinBERT against FinBERT with customized vocabulary

is a pre-trained model specialized on financial domain, does not outperform a generic PLM such as RoBERTa.

RQ3: To what extent does the vocabulary impact the performance of pre-trained language models in the multi-class text classification task?

Prior works have established that the vocabulary can have an impact on the performance of the pre-trained language model BERT. For example, SciBERT [7], which includes the custom SciVocab vocabulary, achieved better results than the BERT base model for scientific papers. Although SciBERT was trained on Biomedical and Computer Science papers from scratch, we note that its custom vocabulary has only a 42% overlap with the BERT vocabulary, suggesting that the custom vocabulary contributed to the performance improvement.

We propose to investigate the overlap between the vocabulary of FinBERT, which is the same as for BERT, and the actual vocabulary of documents extracted from *Proprietary-1* and *Proprietary-2* datasets of our experiments. Hereafter we refer to the latter vocabulary as our custom vocab[6]. The overlap of our custom vocab with BERT Vocab is 15%, which is even lower than the overlap found for SciBERT. This finding suggests that changing the vocabulary of FinBERT could be required to achieve the improved performances over the BERT base model that was initially expected.

Therefore, we have employed our custom vocab[7] with the FinBERT model. Table 6 describes the performance that is obtained with FinBERT-custom (i.e., FinBERT with our custom vocab) against the performance of FinBERT (i.e., with the BERT vocab). We note that, while the small overlap between vocabularies suggested the custom vocab could lead to some performance improvement, the results do not meet this expectation. These results suggest that the performance of PLMs cannot be simply increased by adapting the vocabulary. Instead, a full pre-training from scratch may be necessary to actually take advantage of the custom vocabulary as well as the specificities of the training dataset.

RQ3 Answer: Using a custom vocabulary on the pre-trained FinBERT model does not appear to be sufficient for yielding higher performance than what could be obtained with the BERT generic vocabulary for the classification task of financial documents.

5.2 Threats to Validity

Our empirical study carries a few threats to validity, which we have attempted to mitigate. First, the general insights that we provide in the application of PLMs may not generalize beyond the specific task of full-text multi-classification. It was, however, the focus of this study. Secondly, the financial documents that form our dataset come exclusively from our industry partners and may thus be very specific. However, this dataset is of significant size and is associated to real transactions with clients from around the world. Unfortunately, at this point, we cannot share these proprietary documents due to legal constraints. Finally, we have investigated FinBERT as a recent approach to specializing a pre-trained BERT model. Although it is not yet considered as a state of the art in the literature, it is the most relevant work we have found in the literature, and the intuition behind its re-training appeared relevant for our investigations.

[6]The custom vocabulary is generated by SentencePiece on the union of *Proprietary-1* and *Proprietary-2*. SentencePiece is available at https://github.com/google/sentencepiece

[7]Accordingly, we resized the FinBERT model to fit the size of our vocabulary

6 CONCLUSION

In this study, we investigated multi-class text classification task in the finance domain. We assessed the performance of several generic PLMs on public generic datasets as well as on proprietary datasets of real-world financial documents. We then assessed the added value of FinBERT, which is a PLM tailored to the financial domain. However, we found that FinBERT was unable to obtain higher performance than the generic PLMs on our financial document classification task. We investigated whether a custom vocabulary could improve the performance of FinBERT. Our experiments show that it did not.

Overall, while the performance that we obtained on our *Proprietary-2* dataset is sufficiently high to consider an integration into financial institution's business processes, none of the PLMs we investigated—not even the finance-specialized FinBERT—achieved sufficient accuracy on our *Proprietary-1* dataset to be put in production.

ACKNOWLEDGMENTS

This work is supported by the Luxembourg National Research Fund (FNR) under the project ExLiFT (13778825).

REFERENCES

[1] 20News. 2008. 20News. http://qwone.com/~jason/20Newsgroups/. Online; accessed January 2021.
[2] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. DocBERT: BERT for Document Classification. *CoRR* abs/1904.08398 (2019).
[3] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In *Mining text data*. Springer, 163–222.
[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 72–78. https://doi.org/10.18653/v1/W19-1909
[5] Chaitanya Anne, Avdesh Mishra, Tamjidul Hoque, and Shengru Tu. 2018. Multi-class patent document classification. *Artif. Intell. Research* 7 (2018), 1–14.
[6] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063* (2019).

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*. 3606–3611.

[8] Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*. 7057–7067.

[9] Matthias Damaschk, Tillmann Dönicke, and Florian Lux. 2019. Multiclass Text Classification on Unbalanced, Sparse and Noisy Data. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*. Linköping University Electronic Press, Turku, Finland, 58–65. https://www.aclweb.org/anthology/W19-6207

[10] 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*. 4171–4186.

[11] Emad Elwany, Dave Moore, and Gaurav Oberoi. 2019. BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding. *CoRR* abs/1911.00473 (2019). arXiv:1911.00473 http://arxiv.org/abs/1911.00473

[12] Eibe Frank and Remco R. Bouckaert. 2006. Naive Bayes for Text Classification with Unbalanced Classes. In *Knowledge Discovery in Databases: PKDD 2006*, Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 503–510.

[13] Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*. 377–384.

[14] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).

[15] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding Convolutional Neural Networks for Text Classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 56–65. https://doi.org/10.18653/v1/W18-5408

[16] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung-Hyon Myaeng. 2006. Some Effective Techniques for Naive Bayes Text Classification. *Knowledge and Data Engineering, IEEE Transactions on* 18 (12 2006), 1457–1466. https://doi.org/10.1109/TKDE.2006.180

[17] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information* 10, 4 (2019), 150.

[18] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

[19] Jieh-Sheng Lee and Jieh Hsiang. 2019. PatentBERT: Patent Classification with Fine-Tuning a pre-trained BERT Model. *CoRR* abs/1906.02124 (2019). arXiv:1906.02124 http://arxiv.org/abs/1906.02124

[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[21] Baoli Li and Carl Vogel. 2010. Improving Multiclass Text Classification with Error-Correcting Output Coding and Sub-class Partitions. In *Advances in Artificial Intelligence*, Atefeh Farzindar and Vlado Kešelj (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 4–15.

[22] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study. *JMIR medical informatics* 7, 3 (2019), e14830.

[23] Clavance Lim. 2019. *An Evaluation of Machine Learning Approaches to Natural Language Processing for Legal Text Classification*. Master's thesis. Imperial College London.

[24] 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* (2019).

[25] Ronny Luss and Alexandre d'Aspremont. 2015. Predicting abnormal returns from news using text classification. *Quantitative Finance* 15, 6 (2015), 999–1012.

[26] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796.

[27] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, Vol. 752. Citeseer, 41–48.

[28] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3428–3448. https://doi.org/10.18653/v1/P19-1334

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[30] Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2019. Enhancing BERT for Lexical Normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, Hong Kong, China, 297–306. https://doi.org/10.18653/v1/D19-5539

[31] Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4658–4664. https://doi.org/10.18653/v1/P19-1459

[32] Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online, 205–214. https://doi.org/10.18653/v1/2020.bionlp-1.22

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[34] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI* (2018).

[35] Jason DM Rennie and Ryan Rifkin. 2001. Improving multiclass text classification with the support vector machine. *Technical report, AIM-2001-026.2001* (2001).

[36] Jason D. M. Rennie. 1999. *Improving Multi-class Text Classification with Naive Bayes*. Master's thesis. Massachusetts Institute of Technology. http://qwone.com/~jason/papers/sm-thesis.pdf

[37] Reuters-TRC2. 2004. Thomson Reuters Text Research Collection. https://trec.nist.gov/data/reuters/reuters.html. Online; accessed January 2021.

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2) co-located with the Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)*. 1–5.

[39] Timo Schick and Hinrich Schütze. 2020. Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 8766–8774. https://aaai.org/ojs/index.php/AAAI/article/view/6403

[40] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/p19-1355

[41] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification?. In *China National Conference on Chinese Computational Linguistics*. Springer, Springer International Publishing, 194–206.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[43] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[45] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[46] Baoxun Xu, Xiufeng Guo, Yunming Ye, and Jiefeng Cheng. 2012. An Improved Random Forest Classifier for Text Categorization. *JCP* 7 (2012), 2913–2920.

[47] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.

[48] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[49] Chin Man Yeung. 2019. *Effects of inserting domain vocabulary and fine-tuning BERT for German legal language.* Master's thesis. University of Twente. http://essay.utwente.nl/80128/

[50] Shanshan Yu, Jindian Su, and Da Luo. 2019. Improving BERT-based text classification with auxiliary sentence and domain knowledge. *IEEE Access* 7 (2019), 176600–176612.

[51] Matt Zames. 2016. 2016 Letter To JP Morgan Shareholders, 2016 Annual Report. https://www.jpmorganchase.com/corporate/investor-relations/document/ar2016-lettertoshareholders.pdf.

[52] H. Zhang and D. Li. 2007. Naïve Bayes Text Classifier. In *2007 IEEE International Conference on Granular Computing (GRC 2007)*. 708–708.