Development and Evaluation of a German Language Model for the Financial Domain

Nata Kozaeva, Serhii Hamotskyi, Christian Hänig
Anhalt University of Applied Sciences
Bernburger Str. 55, 06366 Köthen, Germany
nata.kozaeva@hs-anhalt.de, serhii.hamotskyi@hs-anhalt.de, christian.haenig@hs-anhalt.de

Abstract

Recent advancements in self-supervised pre-training of Language Models (LMs) have significantly improved their performance across a wide range of Natural Language Processing (NLP) tasks. Yet, the adaptation of these models to specialized domains remains a critical endeavor, as it enables the models to grasp domain-specific nuances, terminology, and patterns more effectively, thereby enhancing their utility in specialized contexts. This paper presents an in-depth investigation into the training and fine-tuning of German language models specifically for the financial sector. We construct various datasets for training and fine-tuning to examine the impact of different data construction strategies on the models' performance. Our study provides detailed insights into essential pre-processing steps, including text extraction from PDF documents and language identification, to evaluate their influence on the performance of the language models. Addressing the scarcity of resources in the German financial domain, we also introduce a German Text Classification benchmark dataset, aimed at fostering further research and development in this area. The performance of the trained models is evaluated on two domain-specific tasks, demonstrating that fine-tuning with domain-specific data improves model outcomes, even with limited amounts of domain-specific data.

Keywords: Language Modelling, Financial NLP, German

1. Introduction

In the rapidly evolving financial sector, where precision and accuracy of information dissemination are paramount, the development of specialized Language Models (LMs) becomes not just beneficial but essential. The financial domain is characterized by its dynamic nature, requiring the processing of vast quantities of data that include market reports, regulatory filings, and financial news. Each of these data types is imbued with complex jargon, numerical information, and nuanced expressions specific to the financial industry. The application of specialized language models in this sector enables several promising use cases, including automatic checking for eligibility criteria (Hänig et al., 2023), facilitating automatic financial reporting (Oyewole et al., 2024), and ensuring automatic consistency checking (Ali et al., 2023).

The predominance of English in the global financial literature has led to a wealth of text data in English, ranging from publicly accessible 10-K forms[1] and earnings call transcripts to comprehensive resources like Seeking Alpha[2] and the System for Electronic Document Analysis and Retrieval[3]. In stark contrast, the German financial sector faces a significant challenge due to the scarcity of equivalent resources in the German language, highlighting a critical gap in both financial text corpora and annotated datasets within this domain.

This research addresses this gap by development and evaluation of a German LM fine-tuned for the financial sector. We compare its performance on downstream tasks against a general-purpose German LM (referred to as vanilla LM). Our goal is to ascertain whether a domain-specific LM can surpass the vanilla model in the nuanced task of processing German financial texts. Through a series of experiments involving both the further pre-training of existing LMs and the training of new models from scratch using various dataset configurations, we explore this question in depth.

Research, such as that by Hänig et al. (2023), demonstrates that an English FinBERT model (Yang et al., 2020) fine-tuned for the financial domain falls short in performance when applied to German financial data, compared to a general German LM, which, in turn, outperforms an English model on out-of-domain German tasks.

Considering the features of financial language, including complex sentence structures, formal tone, specialized vocabulary, and legal terminology, the development of a dedicated German LM for the financial domain is imperative. To facilitate the development of German financial LMs, we perform thorough analyses of financial text corpus compilation and study the effect of various pre-processing steps. Furthermore, we create and publish a new German benchmark dataset for evaluation language models in the financial domain.

Our research utilizes the BERT architecture (Devlin et al., 2019), specifically German BERT (Chan

[1] https://www.sec.gov/
[2] https://seekingalpha.com/
[3] https://www.sedarplus.ca/

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 40–49

et al., 2020), drawing inspiration from its application in related fields, including FinBERT (Yang et al., 2020), SciBERT (Beltagy et al., 2019), ClinicalBERT (Huang et al., 2020), and BioBERT (Lee et al., 2019).

1.1. Related Work

The same approach was used to develop models for other domains: ClinicalBERT pretrained on clinical notes (Huang et al., 2020), SciBERT pretrained on scientific papers (Beltagy et al., 2019).

There is a significant shortage of publicly available text corpora and labeled datasets related to financial topics in the German language. The CODE ALLTAG corpus (Krieg-Holz et al., 2016) is a text dataset comprised of emails in the German language. Within this corpus, there is a "FINANCE" collection, which includes 174,375 emails, containing nearly 2.5 million sentences. The Bundesstelle for Open Data has released deutschland[4] and handelsregister[5] to enable the retrieval and download of data from the Bundesanzeiger and Handelsregister, respectively. Data extracted from the Bundesanzeiger has been used in academic research, serving various purposes, such as company name recognition (Loster et al., 2017) and the training of language models on text resembling financial content (Biesner et al., 2022). However, these datasets were not made publicly available.

Jørgensen et al. (2023) conducted a comprehensive analysis of labeled datasets in the financial domain revealing that the vast majority of resources is in English. Only few non-English datasets exist with just one multilingual dataset containing the German language: SIXX-Corpora (Gaillat et al., 2018) for sentiment analysis (non-open dataset).

1.2. Contribution

Our first contribution involves the creation of a German financial dataset suitable for multiclass and multilabel classification tasks. For this we used the MultiFin dataset and translated it in German.

Our second contribution includes development and evaluation of domain-specific LMs for German financial language and thorough analysis of the impact of decisions made during dataset construction and pre-processing on the models' performances.

2. Financial Data for Language Model Training

Delimiting the scope of financial language is challenging, covering diverse subdomains like capital markets, banks, and insurance, with data from varied sources including financial documents, laws, and news. These sources, while thematically aligned, differ in vocabulary and complexity—news articles are generally more accessible, while documents like prospectuses feature domain-specific jargon. Some texts, such as annual reports, follow strict standards, contributing to their uniformity.

Given this linguistic diversity and the specific characteristics of various document types, we opted to construct a dataset that encompasses multiple categories of documents. This approach aims to maximize the dataset's diversity, thereby providing a comprehensive foundation for training and evaluating our language models.

2.1. Financial Document Collection

In this study, we utilize FinCorpus-DE10k (Anonymous, 2024), a domain-specific dataset composed of various document types, as a foundation for our analysis. It features the following document types:

Base and Final Terms Prospectuses Financial prospectuses that provide terms and conditions of the issuance of financial securities. The structure, content, release procedure are regulated by Article 8 and 10 of REGULATION (EU) 2017/1129 ("Prospectus Regulation").

Annual Reports of the Bundesbank Documents providing information about economic and financial issues, monetary policy, risks of financial stability etc. Annual reports usually contain a larger number of data visualizations and images.

International Financial Reporting Standards EU International Financial Reporting Standards (IFRS)[6] from the years 2017–2023. These documents define standards as accounting rules that facilitate understanding and comparability of financial statements across borders to ensure corporate transparency.

Law Documents containing German laws in the financial domain. The core regulations applicable to the financial sector in Germany are laid down in the Banking Act (KWG)[7]; the Securities Institutions Act (WpIG)[8], the Securities Trading Act (WpHG)[9] etc. as well as EU Directives implemented into German law.

Informational Materials Brochures and advertisements in the area of finance, description of financial products and general terms and conditions. Most documents of this collection have a wider variety of fonts, photos, colors, and are mostly aimed at a more general audience.

[4] https://github.com/bundesAPI/deutschland

[5] https://github.com/bundesAPI/handelsregister

[6] https://www.ifrs.org/

[7] https://www.gesetze-im-internet.de/kredwg/

[8] https://www.gesetze-im-internet.de/wpig/

[9] https://www.gesetze-im-internet.de/wphg/

	num txt doc	num tokens	num numeric tok	num sent	mean length tok.	mean length sent.
Final terms	10,986	112,344,212	5,307,180	4,026,251	6	26
Base prospectuses	731	49,353,187	1,996,865	1,435,924	5	28
Annual reports	88	7,406,590	731,624	318,683	6	21
Informational materials	139	2,200,884	68,976	94,071	6	20
Law	138	4,062,628	373,439	95288	6	28
IFRS	7	3,726,002	135,215	107,577	6	30
BBK monthly	412	48,182,195	21,720,392	1,750,691	3	25
News	20	2,144,970	52,497	94,888	6	19
Wikipedia	1	9,181,311	331,821	457,495	6	17
Total	12,516	238,601,979	30,718,009	8,380,868	-	-

Table 1: Document statistics in TXT files

Bundesbank Monthly Reports The initial collection (PDF documents) contains 866 monthly reports of the German Bundesbank from the years 1949–2022.

Statistics of the dataset are provided in Table 1.

2.2. Layout and Text Extraction

The PDF documents contain files with very diverse layouts. Financial information is oft presented using tables and charts, incorporating a large number of figures compared to regular language. Another layout features are columns and table-like structures. The presence of columns and tables can disrupt the linear extraction (corresponds to the natural reading order) of text. In context of pre-training a LM this is important, because otherwise, the attention mechanism will be applied on a sequence with an incorrect token order.

For the experiments pdfplumber [10] library was employed to extract the text from PDF files. Given the uncertainty in document layouts, our initial experiment used a text extraction library without adjustments for specific structures.

Next we conducted text extraction taking into account possible layout differences. Assuming that the document collections likely contain columns and tables due to their financial nature, the impact of an alternative text extraction method on the Model's performance was assessed.

PyMuPDF[11] was used for layout-specific extraction. Upon comparing the results with those obtained using pdfplumber, this solution demonstrated accuracy within the randomly selected documents chosen for comparison. The extracted text was observed in its natural reading order.

2.3. Language Detection

To train a German language model, a critical step is to analyze the linguistic composition of our dataset to ascertain the prevalence and distribution of languages within it. This analysis leverages insights from Anonymous (2024), wherein the authors utilized the automatic language identification tool lingua-py[12] to quantify the language proportions across the document collection.

Within the dataset, the predominant language is German, succeeded by English, while the presence of other languages is comparatively minimal. It is presumed that the detection of other languages originates from language detection inaccuracy. Figure 1 illustrates a histogram of language distribution within the dataset, denoting German, English, and other languages.

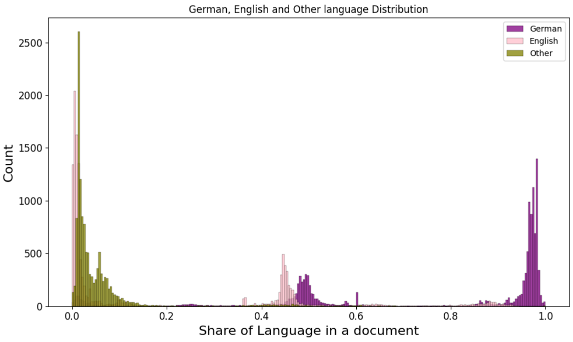

Figure 1: Language distribution for each document of the corpus

The dataset predominantly features documents exclusively in German. There is also a significant subset of bilingual documents, with German and English content, primarily between 40-60%. A smaller fraction of documents includes *Other* languages paired with either English or German. Trilingual documents, which are relatively scarce, are likely artifacts of language identification errors and are considered noise. To refine the dataset for German language specificity, the language detection algorithm from SpaCy[13] was employed to segregate and remove English language texts, thereby curating a corpus composed solely of German language.

[10] https://github.com/jsvine/pdfplumber

[11] https://github.com/pymupdf/PyMuPDF

[12] https://github.com/pemistahl/lingua-py

[13] https://spacy.io/

2.4. Corpus Compilation

For the experiments, the financial data was augmented with common language data, utilizing the *Wortschatz* collection (Goldhahn et al., 2012) from Leipzig University[14] to create a corpus of common German language. This corpus consists of separate sentences of varying length. In the process of training a LM, each sentence serves as an individual instance or training example. Given that the financial corpus is composed of documents, it inherently contains more contextual information compared to isolated sentences. Consequently, at this point, the aim was to incorporate an additional common language corpus that comprises full texts rather than discrete sentences. The German colossal, cleaned Common Crawl corpus[15] was employed, comprising texts of varying lengths.

Further the results of LM performance for mixed datasets (financial corpus mixed with common language sentences and financial corpus mixed with common language texts) will be compared. The token count in both corpora of common German language is approximately equivalent to that of the financial corpus so that the token count in the mixed corpora is approximately double that of the financial corpus.

At this point term frequency was calculated and sorted in the financial and common language corpus to check to which extent the domain specific dataset vocabulary varies from the common language. There was a considerable contrast between the two corpora, emphasizing the financial corpus's domain-centric nature.

2.5. Corpus Configurations

By employing two text extraction methods, language detection and mixed corpus, we analyzed an array of data combinations.

From a *data* perspective, the following preprocessing configurations were explored:

none Text is extracted as it is.

language detection German-only language extraction (leveraging language detection).

layout detection Extraction accounts for document layout (applying columns and tables detection).

layout & language detection Extraction considering both layout detection and German-only language extraction.

From a *domain-focused* perspective, examination encompassed:

fin Financial data is used.

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

Table 2: Overview of High-Level tags across the 6 classes used in the multiclass classification task (Jørgensen et al., 2023)

.

mixed A combination of financial and general language data is used.

Regarding *mixed* data, the following data is added to the financial corpus:

sentence General German language sentences.

text General German language text providing a larger context.

text and sentence Both sentence and text data.

3. Financial Datasets for Downstream Evaluation

In the context of financial language processing, the evaluation of language models on domain-specific tasks is crucial for assessing their practical utility and effectiveness. This section delves into the use of two pivotal downstream tasks: Text Classification (TC) and Named Entity Recognition (NER), which serve as benchmarks for evaluating the performance of our fine-tuned German financial language models.

3.1. Financial Text Classification Dataset

Text Classification in the financial domain involves categorizing text into predefined categories, an essential function for organizing and interpreting vast amounts of financial data. Our new benchmark dataset is based on the MultiFin dataset (Jørgensen et al., 2023), a rich collection of real-world financial article headlines annotated with both high-level and low-level topics. The original MultiFin dataset consists of 10,048 real-world financial article headlines in 15 languages. The dataset is annotated with 6 high-level topics and 23 low-level topics for multi-class and multi-label classification, respectively (see Table 2, Figure 3). For the multi-label classification task, there are up to 3 annotations per example, which sums up to 14,230 annotations with an average of 1.4 annotations per example.

OpenAI API gpt-3.5-turbo[16] was used to translate the dataset examples from the source languages to German. Each example was accompanied by a specific prompt that included the source language

[14] https://corpora.uni-leipzig.de

[15] https://german-nlp-group.github.io/projects/gc4-corpus.html

[16] https://platform.openai.com/docs/models/gpt-3-5

from the dataset. This guided the model more effectively, eliminating the need for language detection as the source language was explicitly provided.

Given the dataset's multilingual nature and time constraints, exhaustive manual verification of each translation was impractical, making it impossible to guarantee translation perfection. To evaluate translation quality, we selectively reviewed 100-150 examples per class across English, Italian, and Russian, focusing primarily on semantic accuracy. Translations were classified as either *semantically correct* or *semantically incorrect*, with the latter category excluded from further grammatical evaluation due to their failure in meaning transmission. This methodology confirmed that the translations maintain a quality level adequate for their intended analytical use, as evidenced by the outcomes illustrated in Figure 2:

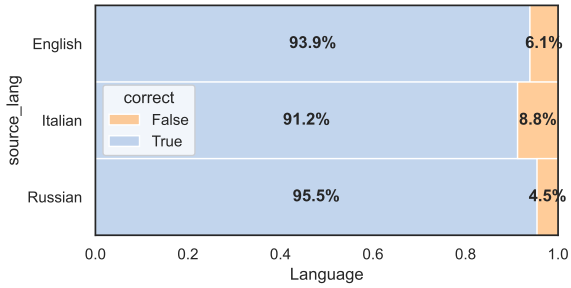

Figure 2: Language distribution for each document of the corpus

The original MultiFin dataset comprises three subsets: train, dev, and test, containing *6430*, *1608*, and *2010* examples, respectively. The German MultiFin dataset features the same number of instances per split as the original MultiFin dataset, as all instances have been translated to German.

Given the problem of imbalanced classes (Kubat, 2000), instances for each class in each subset were counted. This was done to ensure that each subset (train, val, test) contains a proportional number of examples for each class (see Figure 3).

The created German MultiFin Dataset is available on HuggingFace[17].

3.2. Financial Named Entity Recognition Dataset

Named Entity Recognition (NER) in the financial domain seeks to identify and classify key information pieces from unstructured text, such as financial instruments, criteria, and terms. For this task, we use a dataset for examination of eligibility criteria from securities prospectuses (Hänig et al., 2023) which is meticulously annotated across 17 distinct entity classes.

Target type	Train	Test
coupon fixed	431	375
coupon variable index	56	84
coupon variable margin	38	42
coupon variable operator	37	43
coupon variable tenor	45	75
currency	514	577
early redemption amount	64	52
early redemption	177	108
isin	421	417
principal amount	784	800
redemption at maturity amount	26	42
redemption at maturity	370	347
special termination	96	109
special termination amount	61	63
status non preferred	56	47
status senior non preferred	488	333
type of instrument	431	422

Table 3: Number of annotations per target type in the dataset splits (Hänig et al., 2023).

Being able to detect this array of classes empowers models to advance the automation process for determining the eligibility criteria of securities prospectuses issued by central banks, addressing eight intricately varied criteria essential for evaluating an issuance's eligibility. The criteria encompass a broad spectrum, including:

- Coupon
- Currency
- Early redemption amount
- Principal amount
- Redemption (amount) at maturity
- Special termination right
- Liquidation Status (Senior/Subordinated)
- Type of instrument

The documents were annotated manually, to assess consistency of the manual annotation process the authors measured inter-annotator agreement (IAA) using Intersection over Union (IoU). The resulting IAA scores range from 0.731 to 0.932 (Hänig et al., 2023). The total number of annotations per type are shown in Table 3. The annotated data was converted and transformed into a dataset for token classification, namely into BIO-encoded sequences. The labels were aligned to the tokenization of the BERT model.

4. Language Model Training

32 distinct training experiments were conducted, categorized based on various factors, which were to be explored.

The factors encompassed aspects described in subsection 2.5. Each data aspect was used for different model *weight initialization*:

pre-trained model training uses pre-trained weights,

from scratch model training uses randomly initialized weights.

[17] https://huggingface.co/datasets/anhaltai/german-financial-dataset

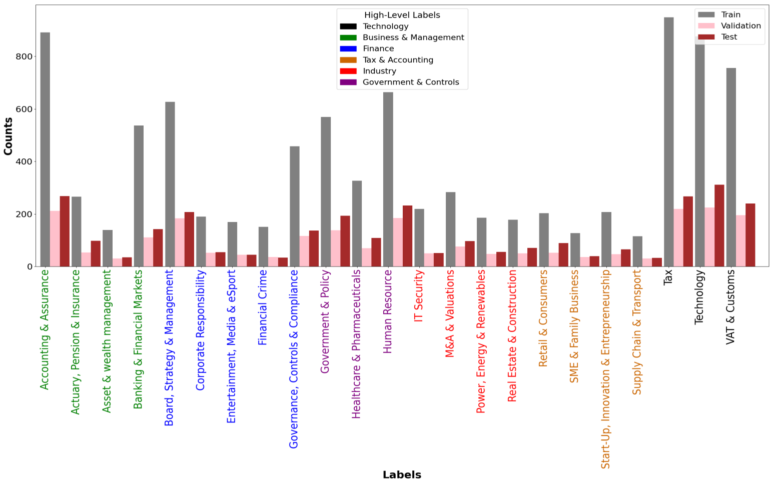

Figure 3: Distribution of Labels Across Training, Validation, and Test Sets. (The bars represent the distribution of low-level labels, with colors corresponding to high-level labels.)

4.1. Training Results of Language Models

For language model training, we report loss scores which directly correspond to the commonly used intrinsic language model evaluation metric Perplexity. The following regularities can be observed (see Table 4):

Comparing models based on *weight initialization*, pre-trained models consistently outperform the models initialized from scratch in all experiments.

From a *data* perspective, language detection improves the results for four models , but slightly lowers the performance in the other four compared to models without language detection. *Layout detection* consistently contributes to the model performance.

The results obtained when *layout detection* has been applied outperformed all other models except for the model trained from scratch using a mixed dataset comprising text and sentence examples.

When comparing different example compositions within *mixed* datasets, an evident pattern can be observed. Pre-trained models leveraging text examples tend to outperform other variants (sentences combined with texts or solely sentences). Conversely, models trained from scratch perform better when trained on mixed datasets with sentence compositions.

In the comparison between text & sentences and solely sentences compositions for pre-trained models, the text & sentences approach is worse for the model without layout and language detection, it fares better for the other three models.

The best model performance is achieved when fine-tuning a pre-trained model on financial data using layout detection. This configuration achieves a loss of *0.72*.

5. Evaluation

5.1. Text Classification Task

In *multiclass* classification task two models outperform the baseline (vanilla German BERT-base) model: the model *pre-trained on financial corpus with language detection* and the model *pre-trained with layout and language detection on a mixed dataset with text and sentence composition*. The LM model, that exhibited best result based on intrinsic metrics (cross-entropy loss and perplexity) did not achieve the best score for the downstream task. Conversely, the poorest-performing LM, trained from scratch on financial data without language and layout detection, similarly demonstrated worst performance for the downstream task.

While LM results indicate that models with both layout and language detection consistently achieved inferior results compared to those with layout detection, the downstream task results present a more nuanced picture. Five models incorporating language and layout detection show better performance on multi-class classification and four on multi-label classification compared to those employing layout detection, only.

Among from scratch-trained models, one model stood out with notably lower loss compared to other from-scratch counterparts. This model demonstrates a slightly better performance in this downstream task (0.8537) compared to most scratch-trained models, except for the model featuring layout and language detection trained on mixed data with text example composition (0.8811).

The analysis of the downstream task suggests, that a small loss (or a small perplexity) does not guarantee great performance on this downstream task.

Multi-label classification results are shown in

	fin	fin	mixed data	mixed data	mixed data	mixed data	mixed data	mixed data
			text&sent	text&sent	sent	sent	texts	texts
	pretr	from scr	pretr	from scratch	pretr	from scratch	pretr	from scr
none	1.11	7.33	1.44	**2.78**	1.32	4.51	0.91	**5.48**
language detection	0.92	5.33	1.37	5.03	1.43	4.06	1.02	5.58
layout detection	**0.72**	**5.28**	**1.22**	4.30	**1.28**	**3.89**	**0.83**	**5.48**
layout & language detection	0.79	5.29	1.30	4.90	1.35	4.05	0.91	5.54

Table 4: Comparison of Language Model Training Results (loss values)

	none	language detection	layout detection	layout & language detection
fin pretrained	0.8829	0.8849	0.8915	**0.8957**
fin from scratch	0.0	0.8151	0.8261	0.8201
mixed text & sent pretrained	0.8819	**0.8940**	**0.8934**	0.8842
mixed text & sent from scratch	0.8547	0.8379	0.8901	0.8359
mixed sent pretrained	0.8834	**0.8923**	0.8905	0.8890
mixed sent from scratch	0.8282	0.8361	0.8269	0.8333
mixed text pretrained	**0.8944**	**0.8967**	0.8890	**0.8957**
mixed text from scratch	0.8175	0.8165	0.8329	0.8380

Table 5: Multi-class / multi-label TC results on downstream dataset (macro-averaged F-score)

Table 5. For this task, *seven* models outperform the baseline model, additionally one model (pre-trained with financial data with layout detection) achieves the same results as the baseline.

The baseline model exhibited comparable results for both multi-class and multi-label tasks, with performance metrics of 0.891 and 0.8915, respectively. In contrast, the trained models displayed varying degrees of performance across these tasks.

Among the models that surpassed the baseline in this task, two belong to the *layout and language detection* category, while three models were trained without layout and language detection enabled.

Concerning mixed data, text example composition seems to have a positive impact, as three pre-trained models of this category outperformed the baseline model.

The from scratch trained model using financial data without language and layout detection ceased training after just 2 epochs due to its inability to improve results, yielding a 0.0 F-Score.

The most successful from scratch trained model for this task was the model with layout detection and sent & text example composition (*0.8901*).

Models utilizing *layout detection* generally outperformed those lacking this feature, with one exception observed in the case of a model trained on mixed data using text example compositions. This could be attributed to the substantial dataset providing more contextual information, countering the negative impact of the absence of layout detection.

On the other hand, among models with language and layout detection compared to those with language detection only, three models of the first category outperformed the language detection.

5.2. Named Entity Recognition Task

Results for the NER downstream task are shown in Table 6. The F1-score is calculated separately for every class in the dataset. Additionally, macro-averaged F1-scores are reported to provide a single performance indicator.

Seven models (highlighted with bold font) outperformed the vanilla German BERT base (0.738). All of them belong to the category of pre-trained models while in every pre-trained model category there is at least one that outperformed the vanilla model. Three of them are pre-trained on data with *layout and language detection*. For this downstream task the model *pre-trained only on financial corpus with layout and language detection* achieved the best results. This might be explained by a a strong domain-focus in the data of the NER task.

As outlined in 3.2, the dataset comprises securities prospectuses annotated according to a predefined set of eligibility criteria. The nature of the dataset's content and the specificity of its labels demonstrate a closer alignment with the financial domain than observed in datasets utilized for other downstream tasks. Such alignment enables a comprehensive evaluation of LMs on this dataset to effectively assess their domain-specific performance capabilities.

Augmenting the dataset with common language data with different example composition contribute to the model performance, however the results are slightly worse than of the model pre-trained on the financial data only. Similar to other experiments, models trained from scratch achieve inferior results compared to models using pre-trained weights. Additionally, in contrast to multi-class and multi-label classification, there is a more pronounced disparity in performance between pre-trained models and those built from scratch.

6. Discussion

Our analysis reveals that certain models consistently surpass the baseline German BERT model across all downstream tasks, suggesting that the observed performance gains are systematic rather

	none	language detection	layout detection	layout & language detection
fin pretrained	0.711	0.73	0.732	**0.748**
fin from scratch	0.0	0.387	0.365	0.385
mixed text & sent pretrained	**0.74**	0.695	0.733	**0.74**
mixed text & sent from scratch	0.546	0.501	0.563	0.548
mixed sent pretrained	0.734	0.706	**0.745**	**0.744**
mixed sent from scratch	0.566	0.551	0.59	0.561
mixed text pretrained	0.732	**0.739**	**0.742**	0.724
mixed text from scratch	0.407	0.39	0.4	0.386

Table 6: NER results on downstream dataset (macro-averaged F-score)

than coincidental. This opens up potential for further refinements at both data and model levels.

6.1. Data-Driven Improvements

Models pre-trained on extensive corpora have shown better performance, potentially due to larger data sizes which are critical for models trained from scratch to exhibit comparable results to pre-trained models. For instance, the FinBERT model (Yang et al., 2020) was trained from scratch on sizable corpora exceeding one billion tokens. Similarly, the training dataset for BloombergGPT encompassed a significant token count from financial domains (Wu et al., 2023). This raises the question about the data volume threshold at which models trained from scratch for a specified domain begin to perform on par with pre-trained models.

Deduplication of financial documents presents another research direction, considering the frequent occurrence of redundant text which can affect both training efficiency and cost. Lee et al. (2022)'s work on deduplication indicates potential benefits in training efficiency. However, the impact of deduplication on model perplexity and the balance between content removal and retention of document context has yet to be fully understood. Investigating deduplication at the document level could shed light on its effects.

6.2. Refining Language Detection

Models incorporating both layout and language detection underperformed compared to those utilizing layout detection alone. This discrepancy might be due to the language detection method's word-by-word operation, which can misidentify language transitions in bilingual documents. A sentence-based language detection approach, filtering out sentences with insufficient German content, could preserve context better and improve performance. Assessing this method's impact on both intrinsic metrics and downstream task efficacy is a promising area for exploration.

6.3. Data Filtering Techniques

In datasets like the Bundesbank Monthly Reports, prevalent layout elements such as tables and checkboxes could introduce noise due to a higher ratio of numeric tokens and shorter mean token lengths. Investigating advanced filtering methods or document understanding techniques could be beneficial in addressing these challenges.

6.4. Model and Training Enhancements

Improvements in the training process could include utilizing both Masked Language Modelling and Next Sentence Prediction tasks of BERT for text examples. Further research could explore the impact of training models on additional tasks such as Sentiment Analysis or Named Entity Disambiguation, drawing comparisons with models like BloombergGPT.

7. Conclusions and Future Research

The central aim of this research was to develop a language model specialized for the German financial domain.

A financial corpus was meticulously assembled and two domain-specific datasets were assembled and used for downstream evaluation. The corpus compilation was subject to a series of preprocessing steps and was enriched with a general language data. Furthermore, we created and published the new German dataset *German MultiFin* useful for multi-class multi-label classification in the financial domain.

Across three downstream tasks – multi-class classification, multi-class multi-label classification and Named Entity Recognition – several models displayed enhanced performance relative to the baseline. Particularly, the model pre-trained on the financial corpus incorporating both layout and language detection emerged as superior, yielding the highest average scores across tasks. The strategic inclusion of layout detection, both in conjunction with and independent of language detection, significantly bolstered the performance of pre-trained models in downstream applications. The expansion of financial data with general language content was advantageous for models trained from scratch.

Future research could delve into further refinements, potentially examining alternative language filtering techniques, data deduplication approaches, and the procurement of more domain-specific data.

8. Bibliographical References

Syed Musharraf Ali, Tobias Deußer, Sebastian Houben, Lars Hillebrand, Tim Metzler, and Rafet Sifa. 2023. Automatic Consistency Checking of Table and Text in Financial Documents. *Proceedings of the Northern Lights Deep Learning Workshop*, 4.

Anonymous Anonymous. 2024. FinCorpus-DE10k: A Corpus for the German Financial Domain. In *Accepted for the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. ArXiv:1903.10676 [cs].

David Biesner, Rajkumar Ramamurthy, Max Lübberinf, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Lotz, Christian Bauckhage, and Rafet Sifa. 2022. Anonymization of German financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, 13(2):151–61.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. ArXiv:2010.10906 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understandng*. Association for Computational Linguistics.

Thomas Gaillat, Manel Zarrouk, André Freitas, and Brian Davis. 2018. The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. ArXiv:1904.05342 [cs].

Christian Hänig, Markus Schlösser, Serhii Hamotskyi, Gent Zambaku, and Janek Blankenburg. 2023. NLP-based Decision Support System for Examination of Eligibility Criteria from Securities Prospectuses at the German Central Bank. In *Proceedings of AAAI23 Bridge 8: AI for Financial Institutions*, Washington, D. C., USA.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A Dataset for Multilingual Financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Ulrike Krieg-Holz, Christian Schuschnig, Franz Matthies, Benjamin Redling, and Udo Hahn. 2016. CodE Alltag: A German-Language E-Mail Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2543–2550, Portorož, Slovenia. European Language Resources Association (ELRA).

Miroslav Kubat. 2000. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Fourteenth International Conference on Machine Learning*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. ArXiv:1901.08746 [cs].

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. ArXiv:2107.06499 [cs].

Michael Loster, Zhe Zuo, Felix Naumann, Oliver Maspfuhl, and Thomas Dirk. 2017. Improving Company Recognition from Unstructured Text by using Dictionaries.

Adedoyin Tolulope Oyewole, Omotayo Bukola Adeoye, Wilhelmina Afua Addy, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and hinonye Esther Ugochukwu. 2024. Automating financial reporting with natural language processing: A review and case analysis. *World Journal of Advanced Research and Reviews*, 21(3):575–589.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. ArXiv:2303.17564 [cs, q-fin].

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. ArXiv:2006.08097 [cs] version: 2.