FinCorpus-DE10k: A Corpus for the German Financial Domain

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

Abstract

We introduce a predominantly German corpus comprising 12.5k PDF documents sourced from the financial domain. The corresponding extracted textual data encompasses more than 165 million tokens derived predominantly from German, and to a lesser extent, bilingual documents. We provide detailed information about the document types included in the corpus, such as final terms, base prospectuses, annual reports, information materials, law documents, international financial reporting standards, and monthly reports from the Bundesbank, accompanied by comprehensive statistical analysis. To our knowledge, it is the first non-email German financial corpus available, and we hope it will fill this gap and foster further research in the financial domain both in the German language and in multilingual contexts.

Keywords: Text Corpus, Financial Domain, German, Bi-lingual

1. Introduction

The study of financial language is pivotal for understanding the intricacies of global economics, legal frameworks, and business communications. In the pursuit of unraveling the complexities of financial discourse, the availability of diverse and comprehensive linguistic resources is paramount. In this context, we present a significant contribution to the field in the form of a German corpus, offering profound insights into the German financial domain. While large language models perform well on many tasks, there scenarios where fine-tuning language models is beneficial compared to employing large language models. One of these scenarios is decreasing the model size to optimize runtime (Biesner et al., 2022). But the performance on specific tasks can benefit as well, e. g. in the clinical or financial domains (Jørgensen et al., 2023). Domain-specific language models achieve higher accuracy for sentiment analysis in the financial domain and English language (FinBERT (Araci, 2019) achieved a 15% improvement in accuracy) and token classification (Biesner et al., 2022).

Financial text is characterized by a unique vocabulary with implications including sentiment analysis, e.g. many words like liability / share / stock / bull having different connotations compared to the general language (Mishev et al., 2020). This phenomenon also exists for other languages than English, still, some languages lack the availability of language resources (e.g. German). Some studies suggest that other languages might benefit from domain-specific corpora and language modes, e.g. Hänig et al. (2023) show for Named Entity Recognition in the financial domain that model accuracy benefits from domain-adjusted language models in a cross-language scenario.

Jørgensen et al. (2023) note the recent effort to produce monolingual financial BERTs to process financial text (while highlighting the need for, and importance of, multilingual financial datasets and models).

We present a corpus consisting of 12.5k financial documents as PDFs (mostly in German, with some documents bilingual - German and English) to stimulate the area of German NLP in the financial domain. To our knowledge, there's only one other corpus of German financial language - the email-based CODE ALLTAG (Krieg-Holz et al., 2016).

Our corpus is composed of seven document types, organized in an intuitive directory structure accompanied by relevant metadata.

Potential uses of our corpus include tasks in the field of Natural Language Processing like Language Model Fine-Tuning (for financial tasks), Document Understanding (e.g. document structure extraction) or OCR (parallel visual and textual data).

2. Related Work

German is traditionally considered a high-resource language and a large amount of both general-purpose and specialized corpora exist and are publicly available. Surprisingly, there's a notable gap in the financial domain. CODE ALLTAG (Krieg-Holz et al., 2016)[1] is a text corpus composed of German-language emails from Usenet groups, and it contains a "FINANCE" collection containing 174,375 emails and almost 2.5M sentences. To our knowledge, this collection is the only German financial corpus that is freely available.

Data from the *Bundesanzeiger*[2] has been used in the literature for similar purposes, e.g. company name recognition (Loster et al., 2017) or training

[1] https://github.com/codealltag/CodEAlltag

[2] https://en.wikipedia.org/wiki/Bundesanzeiger

LREC-COLING 2024, pages 7277–7285
20-25 May, 2024.

language models on data "that is similar to financial text" (Biesner et al., 2022); none of these datasets have been made available.

The *Bundesstelle für Open Data* published two python packages, `deutschland`[3] and `handelsregister`[4] that allow querying and downloading data from the *Bundesanzeiger* and *Handelsregister*[5] respectively, but no static financial corpus exists.

While not belonging to the financial domain, legal corpora in German exist, most notably Open Legal Data's (Ostendorff et al., 2020) dataset of 100k German court decisions and 444k citations[6].

Most work done in the NLP financial community is carried on in a monolingual English setting (Jørgensen et al., 2023), but as the financial environment is multilingual (as evidenced, in part, by the German+English prospectuses in our own corpus) the relevance of multilingual resources increases; see (Jørgensen et al., 2023) for an overview of English, non-English and multilingual financial datasets built for a specific downstream task, most often Named Entity Recognition and text classification.

3. FinCorpus-DE10k Dataset Description

3.1. Dataset summary

The corpus contains **12,235** PDF files of financial documents (mostly security prospectuses) from seven collections, most with less than 100 pages, as well as the corresponding plaintext files for approx. 10,500 of them. The documents are predominantly (71%) in German, the remaining ones are bilingual (German and English).

The basic statistics by collection can be seen in Table 1.

Metadata for the files is provided in `./metadata.csv`. It contains the following columns (with 4-9 empty if no text was extracted):

1. collection: the name of the collection the file belongs to
2. pdf_only: True if the extracted text of the document is not included
3. pdf_fn: the path to the PDF file
4. txt_fn: the path to the .txt file with the extracted text if present
5. num_pages: number of pages present in the PDF
6. num_chars, num_tokens, num_sentences: number of characters, tokens and sentences, respectivelly
7. token_len: mean token length in characters
8. sentence_len: mean sentence length in tokens
9. tokens_per_page: mean number of tokens per PDF page
10. language: languages present in the file, either "DE" or "EN,DE"
11. ISIN, country: only in Final Terms documents with ISIN filenames

Due to the variety of sources included in the dataset, the different sub-collections are released under different licenses. Unless stated otherwise, the license is Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0[7]). The "Bundesbank Monthly Reports" and "Annual Reports" collections are released under the CC Attribution-NonCommercial-*NoDerivs* 4.0 International license (CC BY-NC-ND 4.0[8]). "Informational Materials" and "IFRS" don't have a specific license attached. To ensure making the corpus as widely accessible as possible, it's released in two version. The first, openly available, contains only the collections releasable under Creative Commons licenses[9]. The second - "complete" [10] - contains all collections, including "IFRS" and "Informational Materials", and will be made available on request.

We diligently adhered to the licensing terms to the best of our understanding and in good faith, but the responsibility for the use and compliance with the applicable law rests upon the final users. In the event that documents included in any of the collections unbeknownst to us have different licenses than the one stated, those licenses shall take precedence. Although extensive efforts were made to identify and exclude such documents, we will promptly remove any documents from the dataset if they are found to be infringing.

The code used for the generation and cleanup of the corpus is available on GitHub[11].

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

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

[5] https://en.wikipedia.org/wiki/German_Commercial_Register

[6] http://openlegaldata.io/research/2019/02/19/court-decision-dataset.html

[7] https://creativecommons.org/licenses/by-nc/4.0/

[8] https://creativecommons.org/licenses/by-nc-nd/4.0/

[9] https://huggingface.co/datasets/anhaltai/fincorpus-de-10k

[10] https://huggingface.co/datasets/anhaltai/fincorpus-de-10k-complete

[11] https://github.com/AnhaltAI/fincorpus-de-10k-scripts

	num pdf	num txt	num pages	mean num pages	num tokens	num sent.	mean length sent.
Final terms	10,450	9,591	222,923	23	100,142,176	3,958,882	25
Base prospectuses	593	590	85,976	146	46,676,196	1,954,417	26
Annual reports	88	87	17,637	203	8,959,269	379,565	24
Informational materials	129	127	2,532	20	993,344	46,413	22
Law	134	134	6,934	52	4,563,746	135,394	34
IFRS	7	7	8,854	1,265	4,234,821	181,959	23
BBK monthly	838	0	110,684	-	-	-	-
TOTAL	**12,239**	**10,536**	**455,540**	**-**	**165,569,552**	**6,656,630**	**-**

Table 1: Statistics by document type collection

3.2. Initial Collection and Normalization

3.2.1. Collection

The core of the FinCorpus-DE10k dataset is composed of more than ten thousand securities prospectuses in PDF format. They were gathered by the German Central Bank (Deutsche Bundesbank) from various sources, such as from the websites of various financial institutions and regulatory bodies, as well as from publicly available databases, and are part of the *Final terms* and *Base prospectuses* collections. They were augmented by a number of separate collections described in details in their individual sections.

3.2.2. Filtering and normalization

Firstly, corrupted, password-protected or otherwise unfit PDF files were filtered out.

The PDF format allows textual information in the form of *PDF text elements*, either added during creation or from e.g. OCR at a later step.

We extracted this text layer with the `PyMuPDF`[12] library, used it for language detection and statistics, and provide it in *txt* files as part of the dataset. It may not fully correspond to the PDF due to OCR and layout considerations, see Section 3.4 for a more comprehensive description of this step.

We used automatic language detection to find and filter out documents in languages other than German or German+English (including the removal of English-only documents), as described in Section 3.5.

We additionally dropped documents that were likely to state issues during further processing, using a number of manual thresholds and heuristics described in Section 3.6.

	pdf	txt
DE	9,501	9,435
XS	255	-
AT	156	156
CH, FR, BE, PT	1-5	-

Table 2: Number of prospectuses by country code (where known)

3.3. Collections

3.3.1. Final Terms Prospectuses

The collection contains **10,450** PDF files, 1 to 719 pages long, with a mean of 25 and a 75th percentile of 32 pages; **98% of files are under 100 pages**.

Files from this and the *Base prospectuses* collections are financial prospectuses that provide terms and conditions of the issuance of financial securities.

95% of the filenames in this collection contain the ISIN (International Securities Identification Number) of the prospectus itself, e.g. "DE000SLB8387.pdf".

An ISIN is composed of three parts: the first two characters are an ISO 3166-1 alpha-2 country code, the next nine numbers are the National Securities Identifying Number (NSIN) that identifies the security, and a single numerical check digit. For securities cleared through Clearstream or Euroclear (which are worldwide) "XS" is used in place of the country code (Röman, 2017).

This allows us to filter them by emitting country: a breakdown can be seen on Table 2. We extracted text only from prospectuses from Germany and Austria.

3.3.2. Base Prospectuses

Base prospectuses contain information about the issuer, description of the security and the summary of the prospectus. This information can be provided as a single document or in three separate ones. The issuer description and the securities note must also include the risk factors spe-

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

cific to the issuer and the security. The prospectuses can be referenced in the final terms documents. The structure, content, release procedure are regulated by Article 8 and 10 of REGULATION (EU) 2017/1129 ("Prospectus Regulation"). The prospectus approval process in Germany is regulated by the The Federal Financial Supervisory Authority (BaFin). Informally, one can see them as the larger documents containing overall information needed for an investor, while the "final terms" documents are issued for each individual security and contain information distinguishing the security, including determining which information from the base prospectus is applicable to it.

Compared to the *Final terms* collection, this collection contains fewer but longer documents; *Final terms* has 16.3 times more documents but only 2.15 times more tokens.

3.3.3. Annual Reports

Contain annual (in a few instances quarterly) reports from (mostly) the Bundesbank and other institutions, spanning the years 1995–2022.

Annual reports of the Bundesbank generally provide information about economic and financial issues, monetary policy, risks of financial stability etc. Annual reports of publicly traded companies consist of standard sections with general corporate information, operating and financial highlights, financial statements, including the balance sheet, income statement, and cash flow statement, Auditor's report etc.

This collection contains a larger number of data visualizations and images.

3.3.4. International Financial Reporting Standards (IFRS)

Contains the EU International Financial Reporting Standards (IFRS) from the years 2017–2023.

These documents describe standards as a set of accounting rules that facilitate understanding and the comparability of companies' financial statements across state boundaries to ensure corporate transparency.

All seven documents are extremely similar, each successive document containing an updated version of the previous one with new additions/deletions.

We want to raise awareness for this duplication within the dataset, because some studies suggest that duplicated data might be detrimental for language model training (Lee et al., 2022). Thorough research experiments need to follow to accurately estimate the impact of duplications in the training data for language model in case of small domain-specific corpora.

3.3.5. Law

Contains files with German laws in the financial and related domains, some in their English translations. The core regulations applicable to the financial sector in Germany are laid down in the Banking Act (KWG); the Securities Institutions Act (WpIG), the Securities Trading Act (WpHG) etc. as well as EU Directives implemented into German law.

It has the longest mean sentence length out of all other collections, likely a reflection of the subject matter as well as more uniform PDF files.

3.3.6. Informational Materials

Contains miscellaneous brochures and advertisements in the area of finance.

They have a wider variety of fonts, photos, colors, and are mostly aimed at a more general audience. In contrast with the *Law* collection, it has the shortest average sentence length of the entire corpus.

3.3.7. Bundesbank Monthly Reports

This collection contains 838 monthly reports of the German Bundesbank from the years 1949–2022[13]. **No extracted text is provided from this collection**, only the PDF documents. The text elements in the documents from the years 1961–1999 (incl.) were absent originally and were added by us.

The dataset is fascinating from a digital humanities standpoint, as a decades-long sequence of documents written in the same context for the same purpose (allowing e.g. to track the changes in the German financial language throughout the years, as well as conventions in the presentation of data etc.).

Some of the reports, especially the older ones, are quite challenging from an OCR perspective (one of the reasons being a large amount of tables and graphs), leading to the quality of both the PDF text layer and the resulting the plain-text representation being one of the lowest of the entire dataset.

We decided to add it to the collection nevertheless, as the stated goal of our corpus is providing a PDF files corpus with German financial language, but without the extracted text, which would have been an outlier by most metrics (e.g. its average token length is almost *half* of the corpus average due to the high amount of OCR artifacts).

3.4. Layout and Text Extraction

The layout of the files isn't uniform and at times relatively complex, precluding the use of trivial layout parsing approaches.

[13] https://www.bundesbank.de/de/publikationen/berichte/monatsberichte

We'll describe in detail on Final terms and Base prospectuses, as they make up the major part of the corpus.

3.4.1. Final terms and Base prospectuses

The *Final terms* and *Base prospectuses* contain, in roughly decreasing frequency of occurrences:

- Columns
- Checkboxes
- Two languages in different configurations
- Table-like structures

The prospectuses were issued by different issuers, with some issuers being represented much more often than others.

Issuers can use different programs to write (or (semi-)automatically generate) the PDF, and have different layout and design conventions.

This has practical implication for potential parsing of these files (be it to extract plaintext suitable for model training or to analyze the prospectuses from a financial standpoint).

As an example, checkboxes are represented in different ways: as character using one of the standard UTF-8 symbols, as character from an embedded font represented with a UTF-8 code point from an Unicode Private Use Area, or as image. All of these can be used in the same document, sometimes - on the same page; sometimes a different strategy is used for checked and unchecked checkboxes (see Figure 3). Detecting/analyzing checkboxes is crucial for automatic evaluation of prospectuses (Hänig et al., 2023), where only valid statements should be considered and invalid statements must be ignored for the eventual analysis.

Similarly, different layouts (esp. columns) lead to difficulties in extracting textual flow consistently from all the documents.

We provide the text we extracted from the documents as plaintext, but it's meant as an approximation and has not been manually checked or corrected. The heterogeneous layout structure of the documents precludes easy text flow extraction. Various tools can use different algorithms to deal with columns, text blocks, tables, and lead to different results. Our choice in that matter is best treated as only one possibility, not necessarily the best one (but fitting for our goal of doing language detection and calculating statistics on document level).

3.4.2. Other collections

Annual and monthly reports contain a very high amount of tables and graphs, but in most ways the points from the last subsection apply.

3.4.3. Manual quality assessment estimation

To evaluate the efficacy of the Optical Character Recognition (OCR) process in addressing known complications within the refined version of the dataset, a methodological examination involving manual spot checks was conducted on a minimal subset of the dataset, comprising 35 documents, with an allocation of six documents per collection, by a duo of annotators[14]. Each document was graded on a scale from 1 to 5 (worst to best), leading to a mean of 4.55, considerably higher than our initial expectations. The primary complications identified were associated with inadequate column parsing, exemplified by the merging of text from distinct columns into a singular plaintext format, and issues pertaining to hyphenation.

Given the constrained sample size, the assessment yields a low-confidence approximation of the OCR process's performance. Nonetheless, it successfully corroborates the existence of two prevalent systematic challenges within the dataset, specifically related to hyphenation and columnar structuring.

3.5. Languages

Most (around 71%) of the text corpus is in German, the remaining documents are bilingual (German and English). The PDF corpus has roughly the same ratios but with lower confidence intervals.While me made an effort to filter out documents in other languages or language combinations, a negligible number of such documents could be present in the corpus.

In the cases where the documents are bilingual, usually it's either a translation (either in two columns or each section/item translated sequentially), or a header block with general info or disclaimers followed by the actual content in another language.

We used the automatic language identification library `lingua-py`[15] and didn't manually verify each of the documents included in the final collection. In particular, we have reason to believe the share of bilingual documents is an overestimation.

In Figure 2 we plotted the distribution of languages in the dataset, reflecting the fraction of each language in each document. The box plot labeled "EN_DE" corresponds to the combined percentage of English and German languages, while the remaining portion, which is ($P_{(OTHER)} = 1 - P_{(EN\_DE)}$), represents all the other languages (labeled as "OTHERS"), or the sum of the percentages of Portuguese, Dutch, French, Nynorsk, Spanish, and Italian (the languages spoken in countries

[14] The raw results are available on Github: https://github.com/AnhaltAI/fincorpus-de-10k-scripts/blob/main/data/humaneval.csv

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

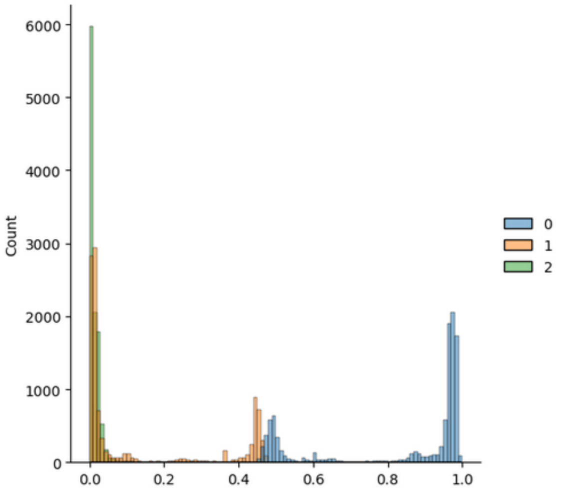

Figure 1: Subsequent to the exclusion of documents containing only English, a graphical representation was constructed to display the three most prevalent languages within each document, organized according to the frequency of occurrence rather than linguistic designation: label 0 signifies the predominant language within the document, label 1 the secondary language, and label 2 the tertiary language. The data reveals two prominent clusters: the majority of documents are monolingual, identified as German (label 0), while a subset exhibits bilingual characteristics, with two languages present in the 40-60% interval (labels 0 & 1). Contrary to our initial assumption that trilingual documents might be observed, languages appearing after the two primary ones represent an insubstantial fraction (label 2), which is typically considered as extraneous noise.

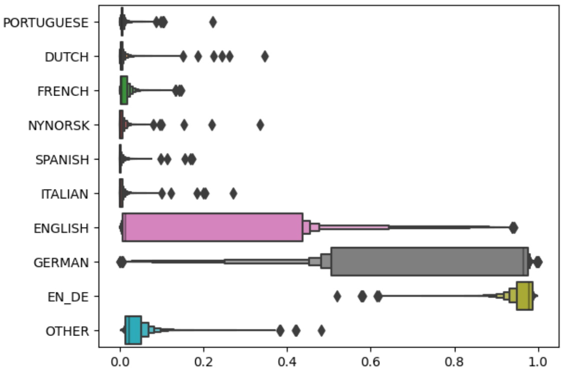

Figure 2: The initial language distribution in the corpus. (EN_DE" is the sum of the English and German percentages in each document, "OTHER" is the sum of all the others ($P_{(OTHER)} = 1 - P_{(EN\_DE)}$))

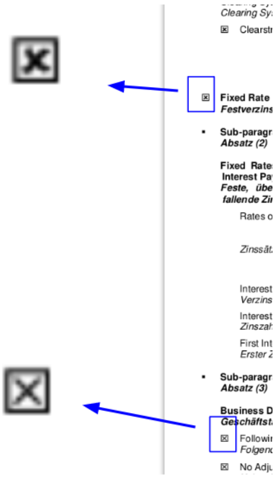

Figure 3: Example of heterogeneous encoding of a checked checkbox in one of the documents

whose country codes was seen in ISINs in *Final terms* documents).

After removing the documents that neither had a high ratio of German nor English from the collection as described in Section 3.6, we converted the values of the languages for each document to the final labels ("DE", "EN,DE") using the following heuristic: if a language had a value over 0.8, the entire document was considered to be in that language (the remaining 20% being mistaken identification), the rest were considered to be bilingual documents.

3.6. Heuristics for detection of noisy documents

During the initial filtering and normalization steps, when manually spot-checking individual documents, we discovered that a number of them are invalid for reasons we hadn't initially anticipated. Manually checking more than 10,000 PDF files was clearly infeasible, and we needed ways to detect and filter out as many of these documents as possible.

We encountered the following failure modes, sometimes in different pages of the same document:

- Encoding issues: when extracting text from the PDF, reading and saving it as UTF-8, some or all of the characters of a document were saved as UTF "REPLACEMENT CHARACTER" U+FFFD (whose function is replacing a character that is unknown or unrepresentable in UTF-8).
- Incomplete text elements in the PDF file: only some of the text in the PDF is machine readable. Scenarios we've seen include:
  - scanned PDF files on which new text is added in a PDF editor, with the new text being readable and the surrounding text being treated as picture

- only one page in the document is readable (multiple PDF files being manipulated on page-level, e.g. replacing/inserting pages in an already existing document)
- only section or numbered list numbers are readable, not the text contained therein

- Bad OCR: The text elements in the PDF were the result of an OCR process resulting in almost unreadable text

- The text elements in the PDF are wrong, either when extracting text or copy+pasting it from an PDF reader. The resulting characters aren't OCR artifacts, they are completely unrelated to the visual ones.

These problems are distinct but were often present in similar (kinds of) documents. We didn't have the goal to fix them, just to recognize with enough recall to filter them out from the dataset.

One option would have been to OCR each file and just use the resulting text, but this could have introduced OCR errors in perfectly machine-readable documents. Comparing the results of the own OCR to the extracted text could point at suspicious documents, a comparison that could be done through string similarity metrics. To detect partial OCR scenarios, comparing the text lengths or the text boundaries on each page could be enough.

We found many such documents among those with unexpected language detection results, and later saw that it's an extremely valuable signal for wrong or partially extracted text: language identification can be challenging in itself, but if the text is small disjoint chunks of text, or mostly numbers and punctuation, the chances of incorrect identifications increase.

Filtering out documents on the basis of language identification results achieves more than one goal and has a high tolerance for errors: removing a document with potentially wrong extracted text is good, but if the extracted text is correct and it's the document itself that's short or atypical, its removal increases the quality of the corpus just as well.

At the end, we found that the following heuristics work for our use case:

- Sum of English and German text is less than 50% of the document

- The entire document contains less than 6 sentences

- The average number of tokens per page is less than 100

- More than 5% of the document contains UTF-8 replacement characters

Lastly, for the Final terms collection, we left only the prospectuses from Germany and Austria, which decreased the number of language detection issues as well as narrowed the list down to documents more likely to be in the two language combinations we needed.

All the documents which were discarded during the process were discarded only from the text corpus but not from the PDF one. Only documents containing exclusively English text were removed completely.

At the end, all this removed 17% of Base prospectuses, 12% of Final terms, 6% of Brochures, 2% of Laws and 1% of Annual reports.

3.7. Availability and Reproducibility

The corpus has been uploaded to Huggingface Hub at https://huggingface.co/datasets/anhaltai/fincorpus-de-10k. The complete version of the dataset (see Section 3.1 on the distinction) will be provided via email application.

The code used to preprocess and analyse the dataset is available at https://github.com/AnhaltAI/fincorpus-de-10k-scripts.

A SemVer-based versioning scheme will be used, with this version being $1.0.0$[16]. We want to preserve the option to update the dataset without breaking any research results that depend on it, with possible changes including the detected language of the document, the addition or removal of documents (e.g. due to licensing issues), and the addition of improved text of the documents. All changes between versions will noted in a changelog.

4. Discussion

4.1. Comparison with CodE AlltagXL

Contrasting with the only other German financial corpus available, CodE AlltagXL's (Krieg-Holz et al., 2016) "FINANCE" collection, one notable difference is that our corpus is built from published material, as opposed to emails.

In the two extremes of language use the authors discuss, on one hand language from high-end performers, conforming to the formal rules of the language (books, manuals, articles, technical papers), and on the other hand - dialogue-oriented, mostly informal and colloquial language from users of different backgrounds (social media, chats, blogs), with the CodE AlltagXL corpus being a mixture of both, our corpus strongly and consistently leans towards the formal side.

[16] https://semver.org/

Some collections contain documents one can consider formulaic, the medium of securities prospectuses especially allows relatively little freedom and variety in the language used. A deeper comparison of language use in both corpora would be an interesting avenue for further research.

4.2. Limitations

As mentioned above, despite the large number of documents, our corpus may have less variety than an user-generated one, though we believe both are equally useful facets of the language used in the financial domain.

One of the important limitations of the corpus is hyphenation on line breaks (especially in documents with multiple narrow columns). We did not handle it explicitly and left it as-is, leading to some words broken into two parts, separated by a hyphen and a newline symbol. One negative impact of this is that this could have inflated the number of tokens in the dataset statistics (see Table 1), but not significantly — the issue is present only in some line endings in some of the documents. The extent to which this impacts the utility of our dataset is debatable. The training of LMs using modern approaches should be stable to the variance introduced by this effect, as they employ their own normalization and sub-word tokenization and therefore rely less on word and line-breaks.

The PDF-first nature of our corpus, along with the complex layouts found in our files, presents a challenge not found in the CodE Alltag (Krieg-Holz et al., 2016) digital-first corpus; parsing complex layouts into a 'natural' text flow is far from being a solved problem. We provide the source PDF files to allow as much freedom as possible in the matter, and in the hope that better approaches will be available in the future.

The bilingual nature of the corpus is, too, affected by this. In some documents the translations are given separately, in some - in parallel columns, in some - each item is given in two translations, one immediately following the other. The use of automatic language detection is as important as layout parsing for the automated plaintext extraction in such PDF files, especially for purposes like building parallel corpora.

5. Conclusions & Future Work

This paper introduces a novel German financial corpus, addressing a significant gap in existing language resources for the financial domain. Currently, only one other German financial corpus is available, specifically focusing on emails. Notably, our corpus includes bilingual documents in English and German, reflecting the growing multilingual nature of the financial sector. Detailed information about the languages present in each document, alongside other metadata, is provided.

Despite the availability of tools like Python packages provided by the German government to download articles from sources such as Bundesanzeiger and Handelsregister, there is an absence of published datasets for building German financial text corpora. This stands in contrast to other domains, such as law, where corpora are readily accessible, and to financial corpora in other languages, particularly English. Surprisingly, even in multilingual financial datasets (as outlined by Jørgensen et al. (2023)), German remains absent.

We hope that our contribution will facilitate the development of similar resources and streamline research efforts built upon them. By providing this corpus, we aim to ease the path for future research endeavors in the multilingual financial domain, fostering a more comprehensive understanding of linguistic patterns within this critical sector.

6. Bibliographical References

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063. ArXiv: 1908.10063 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/journals/corr/abs-1908-10063.bib tex.timestamp: Thu, 29 Aug 2019 16:32:34 +0200.

David Biesner, Rajkumar Ramamurthy, Robin Stenzel, Max Lübbering, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Loitz, 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):151–161.

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).

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 Dirk Thomas. 2017. Improving company recognition from unstructured text by using dictionaries. In *EDBT*, pages 610–619.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. 2020. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*, 8:131662–131682.

Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. Towards an Open Platform for Legal Information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 385–388. ArXiv:2005.13342 [cs].

Jan R. M. Röman. 2017. Trading Financial Instruments. In *Analytical Finance: Volume I*, pages 1–20. Springer International Publishing, Cham.

7. Language Resource References

Krieg-Holz, Ulrike and Schuschnig, Christian and Matthies, Franz and Redling, Benjamin and Hahn, Udo. 2016. *CodE Alltag: A German-Language E-Mail Corpus*. European Language Resources Association (ELRA).