RASMUS KÆR JØRGENSEN

MULTILINGUAL NATURAL LANGUAGE PROCESSING FOR APPLICATIONS IN THE FINANCIAL DOMAIN

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen.

MULTILINGUAL NATURAL LANGUAGE PROCESSING FOR APPLICATIONS IN THE FINANCIAL DOMAIN

RASMUS KÆR JØRGENSEN

UNIVERSITY OF
COPENHAGEN

Doctor of Philosophy (Ph.D.)
Department of Computer Science
Faculty of Science
University of Copenhagen

March 2023

Supervisor:
Christian Igel, University of Copenhagen

Affiliation:
University of Copenhagen
Faculty of Science
Department of Computer Science

ABSTRACT

This dissertation studies natural language processing (NLP) for applications in the financial domain with a focus on multilingual NLP. Together with an industrial partner, three lines of research are pursued.

First, learning systems are devised to automate the accounting task of mapping transactions to accounts, giving a structured overview of a company's finances. The motivation behind this work is that, despite increasing digitisation, much of accounting and bookkeeping is carried out by humans and not machines, and that automatic systems should be able to generalize across different companies. When trained for individual companies, the proposed method processed financial transactions with an accuracy above 80% averaged over 473 companies. Using word embeddings with character-level features to process transaction texts outperformed the baselines of using a lexical bag-of-words representation. After unifying account structures, the system generalized across companies and corporate sectors. A single classifier trained on data from 44 companies belonging to 28 different sectors achieved high performance across companies and corporate sectors, even for unseen companies with no historical data.

The second part studies multilingual domain adaptation and evaluation. Today, most domain-specific models and evaluation datasets are concentrated around English. This motivates studying the benefits of domain adaptive pretraining in a multilingual scenario. The thesis proposes different techniques and strategies for making a single model both domain-specific and multilingual through different compositions of pretraining datasets for continued pretraining of language models, employing adapter-based and full-model pretraining. The results show that the proposed multilingual domain-specific model can outperform the general-domain multilingual model. The single model also performs close to its corresponding monolingual variant. The results hold across different domain-specific datasets representing seven languages and the two pretraining methods. Besides contributing a multilingual financial pretraining corpus and a Danish sentiment dataset to the community, a multilingual financial benchmark dataset covering 15 languages across different writing systems and language families has been created and will be made available.

The third line of research considers the evaluation of the explanations produced by explainability methods for multilingual NLP systems. It is analyzed whether comparable performance figures can be observed across languages or if severe robustness gaps are hidden between related languages. This study is motivated by the need for explainable NLP systems used across languages. The results show, on the provided parallel corpus, that multilingual models perform better on languages seen during fine-tuning, although the unseen languages are part of the pretrained languages. The alignment with human rationales is also better for those languages. However, its also observed that performance on the English language is high, even when not seen during fine-tuning. This suggests that language models favor English and that high accuracy does not necessarily lead to a more successful transfer or a higher alignment with human rationales. The investigation also suggests rank-biased overlap as a suitable metric for rank evaluations and a sequence-wise normalization of LIME's token scores. The study provides a multilingual parallel corpus of rationale annotations in Danish, English, and Italian to benchmark models and explainability methods.

RESUMÉ

Denne afhandling undersøger sprogteknologi (eng. Natural Language Processing, NLP) til anvendelse i det finansielle område med fokus på flersproget NLP. Sammen med en industriel partner bliver tre forskningsretninger undersøgt.

Først udvikles maskinlæringssystemer til at automatisere regnskabsopgaven ved at kontere finansielle transaktioner til konti, hvilket giver et struktureret overblik over en virksomheds økonomi og regnskab. Motivationen bag dette arbejde er, at en stor del af regnskab og bogføring, på trods af stigende digitalisering, udføres af mennesker og ikke maskiner, samt at automatiske systemer skal kunne generalisere på tværs af forskellige virksomheder. Da den ovennævnte metode blev trænet og tilpasset individuelle virksomheder, behandlede den finansielle transaktioner med en nøjagtighed på over 80% i gennemsnit på tværs af 473 virksomheder. Ved brugen af "word embeddings" med n-gram funktionalitet til at behandle transaktionstekster overgik metoden den leksikalske "bag-of-word" repræsentation. Efter at have hamoniseret kontoplanerne generaliserede systemet på tværs af virksomheder og virksomhedssektorer. En enkelt model trænet på data fra 44 virksomheder, repræsenteret blandt 28 forskellige sektorer, opnåede høj ydeevne på tværs af både virksomheder samt virksomhedssektorer, selv for nye virksomheder uden historiske data.

Den anden del afhandlingen undersøger flersproget domænetilpasning og -evaluering. I dag er de fleste domænespecifikke modeller og evalueringsdatasæt koncentreret omkring engelsk. Dette motiverer et studie af fordelene ved domæneadaptiv prætræning i et flersproget scenarie. Afhandlingen foreslår forskellige teknikker og strategier til at skabe en enkelt domænespecifik og flersproget model gennem forskellige sammensætninger af træningsdatasæt til prætræning af sprogmodeller og anvendelse af adapterbaseret og fuldmodel prætræning. Resultaterne viser, at den foreslåede flersprogede domænespecifikke model kan overgå den domænegenerelle flersprogede model. Den enkelte model præsterer også tæt på sin tilsvarende ensprogede variant. Resultaterne gælder på tværs af forskellige domænespecifikke datasæt, der repræsenterer syv sprog, og de to prætræningsmetoder. Udover at bidrage med et flersproget finansielt korpus og et dansk sentimentdatasæt til forskningsmiljøet, er et flersproget finansielt benchmarkdatasæt, der dækker 15 sprog på tværs af forskellige skriftsystemer og sprogfamilier, blevet skabet og vil blive gjort offentlig tilgængeligt.

Den tredje forskningsretning undersøger evalueringen af forklaringerne produceret af fortolkningsmetoder til flersprogede NLP-systemer. Her analyseres, om sammenlignelige effekter kan observeres på tværs af sprog, eller om der er robusthedsforskelle mellem relaterede sprog. Denne undersøgelse er motiveret af behovet for at tolke NLP-systemer, der bruges på tværs af sprog. Resultaterne viser, på det medfølgende parallelle korpus, at flersprogede modeller klarer sig bedre på sprog set under fintræning, selvom de usete sprog er en del af de prætrænede sprog. Enigheden med de menneskelige rationaler er også bedre for disse sprog. Det er derudover værd at bemærke, at ydeevnen på det engelske sprog er høj, selv når det ikke er inkludret under fintræningen. Dette indikerer, at sprogmodeller favoriserer engelsk, og at høj nøjagtighed ikke nødvendigvis medfører en mere vellykket overførsel eller en bedre overensstemmelse med de menneskelig rationaler. Studiet fremlægger også "rank-biased overlap" som en mere passende metrik til rang-evalueringer og en sekvensmæssig normalisering af LIMEs score. Studiet publicerer et flersproget parallelt korpus af rationale annotationer på dansk, engelsk og italiensk til at benchmarke modeller og fortolkningsmetoder.

PUBLICATIONS INCLUDED IN THIS THESIS

This cumulative thesis is based on four peer-reviewed articles. The articles are identical to their original published form with the exception of minor changes in style, format and correction of typographical errors. The following articles are included in the thesis:

[1] Rasmus Kær Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. "MultiFin: A Dataset for Multilingual Financial NLP." In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023.

[2] Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard. "Are Multilingual Sentiment Models Equally Right for the Right Reasons?" In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 131–141. URL: https://aclanthology.org/2022.blackboxnlp-1.11.

[3] Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. "mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3404–3418. DOI: 10.18653/v1/2021.findings-emnlp.290. URL: https://aclanthology.org/2021.findings-emnlp.290.

[4] Rasmus Kær Jørgensen and Christian Igel. "Machine Learning for Financial Transaction Classification across Companies using Character-Level Word Embeddings of Text Fields." In: *Intelligent Systems in Accounting, Finance and Management* 28.3 (2021), pp. 159–172. DOI: https://doi.org/10.1002/isaf.1500. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/isaf.1500. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.1500.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Machine Learning Section at DIKU and PwC for granting me support and funding to pursue a PhD degree.

My sincerest thanks go to my supervisor, Christian Igel, whose never-ending optimism has motivated me to push through difficult times and pursue challenging ideas. Your support has been invaluable: with editing papers, sharing your experience with me, and helping me to understand things that I found difficult, and the list continues. I am grateful that you took me on as a PhD student and continued to have faith in me over the years. This thesis would not be possible without Lars Silberg Hansen, whose endless support and motivation enabled me to complete this PhD. We have had a wonderful journey together that I will always be grateful for and remember.

Additionally, I want to thank all my fantastic colleagues. I genuinely enjoyed spending time with you inside and outside the office. I would also like to thank Desmond Elliott, Mareike Hartmann, and Xiang Dai for being so welcoming and for our collaboration on the two projects. I want to thank Anders Søgaard and Fiammetta Caccavale for working together with me on explainability. Lastly, I want to thank Oliver for the evenings and weekends we spent working together and still sticking together even though things changed around us. Finally, I want to thank my friends and family for reminding me that there are more things to life than the PhD and my girlfriend for all your support, all the love and positivity you shared with me during this process.

I want to express my deepest gratitude to everyone that has in some way or another, helped me to complete this PhD. In the future, when I look back on this journey, I will be reminded of all the wonderful times and people I have been surrounded by.

CONTENTS

Part I

BACKGROUND

1 INTRODUCTION

1.1 PRESENTATION

Financial data is an important part of every company, constantly generated worldwide in high volumes, and in different varieties. Financial text occurs in multiple languages when processing stock market information, tax and accounting data, financial policies or invoices, transactions, and other financial tasks. This makes the financial domain both specific and multilingual.

Financial natural language processing is one of the emerging areas of natural language processing (NLP). It has become integral to the financial domain [8, 129, 138, 176, 217] and gained substantial attention due to the availability of pretrained word embeddings [22, 81, 102], pretrained language models, and the general-purpose transformer models [8, 59, 224]. Recent studies have established that the financial domain benefits from in-domain adaptation [54, 101, 126, 161, 239] and have adapted monolingual models and pretraining resources [8, 54, 122, 126, 239], and released several datasets [50, 92, 99, 138, 176].

However, the financial domain needs automatic systems to accurately process domain-specific data in multiple languages. The advance of the field of financial NLP is challenged by limited and a lack of representative data resources in multiple languages, an absence of multilingual domain adaptive efforts, and a need for stronger benchmarks. To this end, a focus on multilingual NLP is needed to continue the progress in the field of financial NLP.

The objective of this dissertation is to advance multilingual NLP in the financial domain and provide solutions to industrial challenges through research into three areas, developed together with an industrial partner.

FINANCIAL TRANSACTIONS (PART II OF THE THESIS) examines the use of machine learning for the classification of financial transactions across companies using word embeddings with subword information to process transaction texts. This area of investigation was motivated by the fact that accounting and bookkeeping are core financial activities that companies often commission to accounting firms. Accounting firms may benefit from a system that autonomously learns to handle these transactions accurately, even from limited training data. Ideally, it would be a system that learns to classify transactions across companies and corporate sectors, and that is able to generalize to new companies for which little or even no historical data may exist.

MULTILINGUAL FINANCIAL NLP (PART III OF THE THESIS) studies multilingual domain adaptation and evaluation. Domain-adaptive pretraining aims to improve the modeling of text for downstream tasks within a specific domain through continued unsupervised pretraining of a language model on domain-specific text. Motivated by the goal to advance multilingual NLP in the financial domain, this line of research extends the domain-adaptive pretraining to a multilingual scenario. This work also proposes several domain-specific financial resources, including a new financial benchmark covering multiple languages for evaluating multilingual financial language models.

EXPLAINABILITY IN MULTILINGUAL NLP (PART IV OF THE THESIS) investigates the explanations produced by explainability methods for multilingual NLP systems. As NLP systems are deployed, and users interact with these systems, it is important to understand the performance of the employed methods and to which extent humans might align with these methods. Motivated by the need to explain outputs of NLP systems across languages, the study provides a multilingual parallel corpus of rationale annotations in Danish, English, and Italian to benchmark models and explainability methods.

This PhD study was done in collaboration with Pricewaterhouse-Coopers Denmark (PwC)[1] and focused on tackling real-world industry challenges through a research-based approach. PwC is a multinational professional services network that provides services in multiple areas such as accounting, tax, finance, strategy, and technology. PwC was established in the mid-1800s[2] as a London-based accounting firm and has since been a central player in the financial domain. The opportunity for industrial collaboration with PwC has produced interesting results both academically and in practical industrial applications.

1.2 THESIS OUTLINE

The dissertation is structured in six parts. PART I provides the introductory background, including the motivation for the studies, a presentation of the industrial collaboration, and general challenges in the field that shaped the research direction of this thesis.

The original published research is presented in parts II, III and IV. These parts comprise four papers written during the course of this PhD. Each of the papers is presented in a chapter of the thesis, and each represents a self-contained study.

PART II on FINANCIAL TRANSACTIONS examines the use of machine learning for classification of financial transactions across companies. In the paper presented [102], we devise machine learning-based systems that automate the accounting task of mapping transactions to

[1] https://www.pwc.dk/, accessed April 2022

[2] https://www.pwc.com/us/en/about-us/pwc-corporate-history.html, accessed April 2022

accounts – providing a structured overview of a company's finances. The motivation behind this work lies in the observation that, despite digitization, much accounting and bookkeeping is carried out by humans and not machines. Part II also examines some challenges encountered when devising machine learning-based systems considering companies individually and across companies and sectors.

Chapter 4 first investigates a base system that addresses the accounting task for each company individually, that is, mapping transactions to accounts for a single company based on its historic data. Then, in the paper presented [102], we generalize this system to handle transactions across different companies and sectors, even in scenarios where no historic data are available. The capacity to generalize and utilize transfer learning across companies came from the process in which we unified the companies' chart of accounts into a standardized chart. The unification of the output space improves the cold-start problem by making the system applicable to mapping transactions from new, unseen companies with no historic data.

We focus on exploiting the information contained in free-text fields and on deriving features from the other transaction fields. For this, we work on a vector representation of the transaction text and employ different feature engineering methods to arrive at a feature-rich representation. We propose a more suitable method using word embeddings with subword information to process transaction texts that encodes unseen words outside the training corpus and induces better neighborhood similarity. The resulting features derived from the text fields were found to be highly important for classifying transactions to accounts. Our system outperformed the baseline of using a lexical bag-of-words representation. In comparison to rule-based systems, the presented approach offer the possibility of merely retraining the models in order to update the system. Because of our high accuracy and the feedback from accounting experts, we consider machine learning-based systems, like the one presented, as a promising direction in this domain.

PART III on MULTILINGUAL FINANCIAL NLP comprises two chapters and studies evaluation and domain adaptive pretraining, focusing on adapting to multiple languages within a specific domain. This work was motivated by the fact that most existing domain-specific models and evaluation datasets are mainly English-based. This led us to explore the benefits of multilingual domain adaptive pretraining in a single model by extending domain-adaptive pretraining to a multilingual scenario.

Chapter 5 presents the paper "mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model" [101], which studies domain adaptive pretraining of a single model, which can be fine-tuned for tasks within the domain in multiple languages. This enables a language model to both become domain-specific and multilingual. We employ two different continued pretraining methods and investigate different strategies of combining data sources using both mono– and multilingual domain-specific data as well as general-domain data for upsampling. Chapter 5 contributes with a financial pretraining corpus made available for pretraining multilingual financial models and a Danish financial sentiment dataset. For this work, we had the opportunity to partner with other researchers with similar research interests. Together, we started working on multilingual domain adaption in a single model, incorporating the research interest of the two parties involved, namely financial and biomedical applications.

Chapter 6 presents work on a multilingual financial benchmark that covers different writing systems and language families [99]. As in chapter 5, this work contributes to the analysis and comparison of general-purpose and domain-specific models, showing the benefit of domain-adapted models. Apart from analyzing and evaluating models on the benchmark, the primary contribution of the work presented is to promote a more multilingual environment in financial NLP. This benchmark dataset is intended as a resource for developing multilingual financial language models and evaluating how well models can process financial text in multiple languages.

PART IV on EXPLAINABILITY IN MULTILINGUAL NLP finds its motivation in the intersection between the growing body of research in multilingual language models and explainability methods. Although this interest evolved while working in the financial domain, we pursued this work in the general field of NLP building on current explainability research, however, with focus on multilingual settings.

The paper presented in **Chapter** 7 [100] presents an evaluation of explanations produced by explainability methods for multilingual NLP systems. It analyzes whether comparable performance figures can be observed or if severe robustness gaps are hidden between related languages. We propose a trilingual parallel corpus of human rationale annotations for the sentiment analysis task. We also suggest rank-biased overlap as a more suitable metric for rank evaluations as well as a sequence-wise normalization step for LIME's token scores [3]. Using the dataset, we explore popular multilingual models and explainability methods and contribute with experiments in a multilingual setting.

The work in chapter 7 started at the end of my PhD, where I had more freedom to branch out and experiment. I had the opportunity to collaborate on explainability with other researchers who looked into the general domain. The work is based on experimentation to investigate questions that we could not find answered in the literature [100]. The research interest was to look at multilingual scenarios, and we chose to consider a sentiment analysis task using popular methods. Although this is not work conducted on financial texts, the work is still relevant to the objectives of the thesis. The work focused on sentiment analysis, a task of considerable interest to the field of financial NLP [99]. The insights obtained may be helpful to the subset of financial NLP that analyses the sentiment of financial text or general news

[3] A popular explainability tool by Ribeiro, Singh, and Guestrin [187].

media [99]. Similar to the thesis objectives, the work is focused on multilingual NLP and to support more work in multilingual settings.

The dissertation concludes with a discussion in PART V, where the presented studies will be reviewed in the context of the general focus of this thesis, emphasizing natural language processing for applications in the financial domain with a focus on multilingual NLP. At the end of the thesis, the reader finds an appendix VI, which provides technical information supplementing the published papers in parts II to IV.

1.3 MAIN CONTRIBUTIONS

The main contributions of this thesis can be summarized as follows:

i We demonstrated new machine learning systems for supporting accounting firms in mapping financial transactions that generalize across companies and even to new companies, in contrast to the company-specific classifiers or rules used in industrial systems [Ch. 4].

ii We bring forward pretrained word embeddings that account for subword information for mapping transactions to accounts, which alleviated the problem of words being out-of-vocabulary and induced a better neighborhood similarity among the words in transactions. This approach outperformed a baseline BoW approach [Ch. 4].

iii We extended domain adaptive pretraining to a multilingual scenario, achieving the aim of adapting a single model to multiple languages within a specific domain [Ch. 5].

iv We analyzed and compared multilingual domain-adapted models with mono- and multilingual counterparts and found that focusing on multilingual domain-specific methods is a promising direction for future work in the financial domain [Ch. 5 & 6].

v We observed different effects in a multilingual environment by evaluating explanations produced by explainability methods for NLP systems used across languages [Ch. 7]. For example, we found that high performance is not, generally, accompanied by higher alignment with human rationales.

vi We proposed more suitable metrics for comparing ranked rationales and a sequence-wise normalization of token scores [Ch. 7].

During the PhD, we also created several publicly available resources:

i DanFinNews is a Danish financial sentiment dataset for evaluating models on a classification task with domain-specific text.

ii FinMultiCorpus is a multilingual pretraining corpus of financial texts in 14 languages.

iii mDAPT models are publicly available to practitioners and the research community.

iv Multilingual Financial Benchmark is a real-world financial dataset covering 15 languages across different writing systems and language families.

v A trilingual parallel corpus of human rationale annotations in Danish, Italian, and English, for the task of sentiment analysis using the Stanford Sentiment Treebank [209].

These data sets were produced by the authors of the corresponding publications with minor help from participants recruited through our professional network. The annotators were primarily Danish with full proficiency in English. All participated on a voluntary basis, and we have not made use of external paid annotators. We highly appreciate the help that we received in the process of creating these resources for the NLP community.

2 BACKGROUND

This chapter aims to supplement the forthcoming chapters with background on key research milestones and challenges in the field of NLP and financial NLP. The supplementary background provides context behind methods and considerations that have motivated the work behind this thesis.

The introductory background comprises four parts. The first part provides a review on data resources and benchmarks for training and evaluating models. The second part presents representation learning and models in the field of NLP and financial NLP. The third part adds supplementary background to the forthcoming chapter 7 on the explainability of NLP models. The fourth part introduces an overview of industrial collaboration and presents the industrial background relevant to the work behind this thesis. As this PhD has focused on tackling real-world industry challenges through a research-focused approach, chapter 3 also contains considerations for chosen directions and motivations during the PhD.

2.1 DATA RESOURCES AND BENCHMARKS

This section focuses on a important aspect of machine learning research, namely benchmarks and data resources. It first provides general context on benchmarks and data resources, then introduces the field of financial NLP, and lastly, it places the research work of this thesis in context.

Under the *pretrain-then-fine-tune* paradigm [46, 47, 59], this section describes two types of resources commonly used in NLP: BENCHMARK DATASETS, i.e., *labelled datasets* used for test and evaluation, and PRETRAINING DATA, i.e., *unlabelled datasets*, typically used for unsupervised representation learning.

2.1.1 *Data Resources in NLP*

BENCHMARK DATASETS In machine learning research, a benchmark is considered a standard point of reference on which the performance of a machine learning system can be measured and compared across systems. It assists in understanding issues in existing systems and evaluating new approaches [91, 191, 192].

Benchmarks vary with respect to, among other factors, the included tasks, languages, and evaluation metrics applied to measure task performance. A task can consist of one or several combined tasks, and it could represent different challenges or focus on a specific theme, for example, on sentiment benchmarking. In addition, it can be general or specific in its domain, e.g., named-entity recognition (NER) on

Wikipedia or sentiment for stock market prediction. An NLP benchmark can be monolingual or multilingual and contain languages written in different scripts, depending on which tasks and scenarios are targeted by the objective, e.g., NER on English Wikipedia or sentiment benchmark in multiple languages. Another important aspect of a benchmark is how performance should be evaluated, e.g., by calculating accuracy [91], F1 [180], or human agreement [55], and some benchmarks have a summed score across tasks [227]. Besides enabling competition on best performance between evaluated models, a benchmark generally includes a baseline representing a base performance threshold. For this purpose, competitive baseline models are commonly used [202], but some studies also compare against the *human level performance*, which indicates the level of performance that humans reach for a given task [150, 227]. In addition to the benchmark itself, there are supplementary community requirements. Machine learning communites need transparency when benchmarking models against one another, for example, the ability to share and reproduce the results using predefined settings, i.e., fixed splits and experimental setup. Altogether, these properties and requirements help define a benchmark, but many other factors and design principles could play into designing a benchmark [91, 227].

Benchmarks also represent a proxy for the overall development and the current situation as they display the current progress on how far the community has progressed in handling certain tasks. Benchmarks have driven the progress in the machine learning community for decades, with examples such as *MNIST* [118], ImageNet [57], SNLI [23], GLUE [228] and many others. Benchmark datasets continue to evolve in order to challenge research as methods improve. Some examples of improving challenging aspects of benchmarks are GLUE[1][228], XTREME[2][91], and SQuAD[3][181].

The recent advances in multilingual NLP prompted Ruder et al. [192] to extend the multilingual benchmark XTREME [91] to XTREME-R to continue to catalyze progress in the field. Similarly, GLUE [228] is a monolingual benchmark for English and has been extended to SuperGLUE [227] with a new set of more difficult language understanding tasks as new models and methods have driven performance improvements [227]. Also, Rajpurkar et al. increased the question answering dataset, SQuAD, to further challenge the state-of-the-art [180, 181]. One interesting observation is that these benchmarks were extended only a few years after their initial publication. It indicates that machine learning research is advancing at a fast pace and benchmarks need to follow the same pace in order to continue to challenge the community. This makes benchmarks instrumental to machine learning research, and a potential lack of suitable benchmarks could limit further development.

While benchmarks have been widely used to measure progress in machine learning, this approach to measuring such progress has also

[1] General Language Understanding Evaluation (GLUE).

[2] Cross-lingual TRansfer Evaluation of Multilingual Encoders(XTREME).

[3] Stanford Question Answering Dataset (SQuAD).

received criticism. A problem may occur when a research community repeatedly uses the same benchmark for measuring progress. Some benchmarks have been used for nearly a decade [23, 57, 118]. At a certain point, the dataset starts to saturate, and methods begin to overfit the dataset. In addition to overuse, low-quality benchmarks or unchallenging tasks do not reflect the field's current progression [24, 180, 191, 192]. Therefore, a benchmark must be of high quality and reliable [24, 170], which could be ensured by inter-annotator agreement, multiple annotators and annotations, and proper documentation of the dataset, e.g., a DATASHEET [74]. In addition, a benchmark must present a challenging task that not only includes a large number of examples but also concentrates on hard examples. For instance, by including more infrequent and distant languages for multilingual benchmarks [192], by using human curation to create more challenging examples, and by tailoring tasks where models are challenged [24, 180].

PRETRAINING DATA Unsupervised pretraining of language models requires a large-scale unlabelled corpus of text in order to learn a good model of a language, as presented later in section 2.2.1. It is therefore essential to have access to large amounts of text that can be collected and combined into a pretraining dataset, preferably in multiple languages from various sources of high quality since language models ideally should be applicable to any language.

A corpus can include text from books, news, chat forums, articles, *inter-alia*. These texts can be either *general* or *specific*, depending on the particular use of language. General text indicates common language typically found in repositories such as Wikipedia, and specific text refers to source such as biomedical or financial domains that use more domain-specific language. The pretraining dataset can be monolingual or multilingual, that is, a corpus composed of texts in one or multiple languages. The pretraining dataset should ideally represent the language and domain to which the language model will be applied.

Many ways of collecting a pretraining corpus for language representation learning have been suggested in the literature. Popular language models such as BERT and XLM-R are pretrained on pretraining copora collected based on open repositories such as Common Crawl [230] and Wikipedia [22, 59]. However, there are some limiting factors when collecting pretraining data. One important aspect is the availability of text in different languages, especially for low-resource languages [46]. Another is permission to use and access data [101]. Moreover, the fact that models become larger and are trained on more data places a larger demand on training data [43, 46, 124]. Besides the repositories, there are also some published examples of pretraining corpora such as BooksCorpus [248], the English only C4 [178] and its multilingual extension mC4 [237]. In addition, different communities also release data resources helpful for pretraining language models, for example, the PubMed database used for pretraining BioBERT [120] and SEC filings used by Desola, Hanna, and Nonis [58] for pretraining a financial language model.

Despite the success of the massive general-purpose language models trained on massive corpora retrieved from a variety of sources, there are still benefits from pretraining on domain-specific texts [83]. In the next part we focus on the financial domain and the importance of benchmarks and domain-specific text for in-domain performance [8, 101, 161].

2.1.2 *Datasets in Financial NLP*

FINANCIAL BENCHMARKS It is now well established that benchmark datasets are important for machine learning research. The field of financial NLP is interested in a variety of downstream NLP tasks, and more recently, the financial NLP community has presented several new datasets [9, 67, 75, 223].

In chapter 6, we review datasets in the literature and present existing datasets for financial NLP, see Table[4] 1.

(A) Datasets in English		**(B) Non-English datasets**		lang
AnalystTone Dataset [92]	SA	DanFinNews [101] (Thesis, Ch. 5)	SA	DAN
FinTextSen [50]	SA	CorpusFR [95]	NER,RE	FRE
Financial Phrase Bank [138]	SA	BORSAH [6]	SA	ARA
FiQA Dataset [136]	SA,QA			
FinNum-1 [38]	Numeral CLS	**(C) Small multilingual datasets**		
M&A dataset [238]	Deal completeness CLS	ENG-CHI Parallel Fin. Dataset [223]	TC,MT	ENG,CHI
FinNum-2 [37]	Numeral attachment	FNS-2022* Shared Task [67]	SA	ENG,SPA,GRE
StockSen* [235]	SA	SEDAR* [75]	MT	ENG,FRE
FinCausal* [139]	RC,RE	FinSBD-2019* [13]	SBD	ENG,FRE
MultiLing2019 [66]	Summarization	SIXX-Corpora* [73]	SA	ENG,SPA,GER
FIN5 & FIN3 [195]	NER			
Stock-event [119]	Stock Price Prediction	**(D) Large multilingual dataset**		
News-sample OMX Helsinki* [137]	SA	MultiFin (Thesis, Ch. 6)	TC	ENG,DAN,FIN,GRE,HEB,HUN,ISL, ITA,JPN,NOR,POL,RUS,SPA,SWE,TUR
EarningsCall [176]	Stock Price Volatility			
Stocknet [236]	Stock Movement Prediction			

Table 1: A list of datasets for financial NLP with corresponding task (SA=Sentiment Analysis, NER=Named Entity Recognition, QA=Question Answering, TC=Topic Classification, RC=Relation Classification, RE=Relation Extraction, MT=Machine Translation, SBD=Sentence Boundary Detection, CLS=Classification). Marked (*) refers to datasets where a request is needed or an application for permission needs to be obtained before that dataset is shared. Table is from Jørgensen et al. [99] "MultiFin: A Dataset for Multilingual Financial NLP", presented in chapter 6. The orange highlight marks the datasets produced during this PhD.

Most datasets in Table 1 have been published within the last few years and cover essential tasks, such as sentiment and stock market prediction [138, 176], which comprise 12 of 22 datasets. Also, tasks in named entity recognition, relation extraction, and relation classification seem to be pursued by the community [95, 195]. A few studies have considered machine translation and summarization as well as numerical classification [66, 75, 223].

[4] The table is from Jørgensen et al. [99] "MultiFin: A Dataset for Multilingual Financial NLP", presented in chapter 6.

Nevertheless, almost all datasets contain text in English, with only two datasets in non-English languages. There are five multilingual datasets containing more than one language, but only three languages at the most (two trilingual datasets). Altogether, 2/22 are non-English, and 5/22 are multilingual, containing more than one language.

Although there is an interest in evaluating financial models and tasks, also in multiple languages, research in this domain is mainly monolingual, with the focus being primarily on English-centric datasets. This demonstrates a limitation for multilingual NLP, given these datasets' relatively low number of languages compared to the number of languages expected in the real-word tasks. The limitation also hinders financial NLP from adopting new developments or building on multilingual NLP research.

Recent developments in evaluating financial tasks in many languages suggest that there is a growing interest in multilingual NLP. Table 1 shows an increased emphasis since 2020 on non-English and multilingual datasets, and a few studies have also highlighted the need for datasets in other languages [73, 95, 101].

DOMAIN-SPECIFIC PRETRAINING DATA Text varies along several dimensions, and texts considered as *financial texts* can have different sources, e.g., investment reports, financial news and tweets, financial statements, financial regulations, and research papers accepted within the financial topics. Such texts are constantly generated worldwide and occur in multiple languages, making the financial domain inherently multilingual by nature. A corpus for training financial language models must therefore represent these dimensions and, ideally, the multilingual nature of the domain. Although such data might exist on a global scale, it has not been collected and made available to the financial NLP community [99, 101].

Financial NLP is restricted in several respects, making it difficult to obtain large-scale pretraining data from the financial domain [101]. For example, legislation and regulations can make it difficult to collect financial texts, because the financial domain is highly regulated and because financial texts are mainly produced by firms. Another issue arises from confidentiality since financial text may contain sensitive information. Moreover, some firms could have invested in developing resources and intellectual property that they wish to retain internally. A further restriction for financial NLP is that there may simply be no financial texts available for a particular language of interest.

Despite these restrictions, the financial community has several resources, and the literature in the financial NLP field has investigated pretraining datasets, though largely in English [8, 54, 126, 239]. Liu et al. [126] composed an English pretraining dataset from FinancialWeb, YahooFinance and RedditFinanceQA. Also, Desola, Hanna, and Nonis [58] used SEC filings[5], and Araci [8] used RCV1 and TRC2 from the Reuters Corpora [121][6]. Recently Loukas et al. [130] released

[5] http://people.ischool.berkeley.edu/~khanna/fin10-K

[6] https://trec.nist.gov/data/reuters/reuters.html

a large corpus, named EDGAR-CORPUS, of annual reports containing public information about the financial status of companies. The literature seems to contain little on multilingual financial data, besides the RCV2 from Reuters Corpora, which contains texts in more than 10 languages [101].

The size of the pretraining data used in the general domain to pretrain models such as BERT [59], RoBERTa [124], XLM-R [46] is larger in size compared to that used in the financial domain for pretraining FinBERTs [8, 54, 101, 126, 239]. Although more data generally leads to better performance [28, 46, 124], it is also important to obtain diverse text along different dimensions and in multiple languages, as discussed in the previous section 2.1.1.

2.1.3 *General Considerations in the Thesis*

This thesis addresses the gap in data resources and benchmarks important to advancing multilingual NLP in the financial domain. The literature highlights challenges faced by many researchers, namely the lack of unlabelled pretraining data for training domain-specific language models and labeled datasets for benchmarking, preferably in multiple languages.

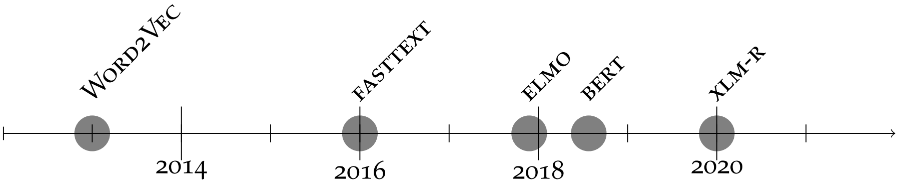
The objective of chapter 5 is to investigate how a single model can become both multilingual and domain specific. This work addresses the shortcomings identified in the literature. In chapter 5, we concentrate on minimizing the lack of multilingual pretraining data for financial texts. While domain-specific data in English seem abundant, other languages have only scarce resources. Therefore, this study tries to address this gap in the data by building a multilingual pretraining corpus based on financial texts.

In terms of existing benchmarks and evaluation datasets in financial NLP, the literature on existing datasets highlights several aspects that need to be addressed and improved to reach the level and extent of large-scale benchmarks such as GLUE (i.e., several tasks) and XTREME (i.e., multiple languages). To address this situation and incentivize research on multilingual NLP in the field of financial NLP, this thesis also focuses on multilingual and non-English resources for evaluation and benchmarking in the financial domain. In chapter 5, we seek to increase the number of non-English datasets available to the financial NLP community. Therefore, we work on a Danish financial news sentiment dataset, as this will allow the financial NLP community to establish combined benchmarks with different properties. For instance, this would allow the combining of several monolingual datasets for a sentiment task or several tasks in several languages in line with the discussion of benchmarks in 2.1.1. In chapter 6, we focus on adding a stronger multilingual property to the existing financial datasets by establishing a multilingual financial benchmark dataset that covers a large number of languages across diverse writing systems and language families. These datasets are also shown in Table 1, marked by a highlight, to depict their relevance in the landscape of existing datasets.

2.2 REPRESENTATION LEARNING IN NLP

This section focuses on representation learning and models. The subsequent segment provides a brief overview and supplements the central methods used in this thesis, with background and context to support the material covered in the individual chapters.

This section is divided into three parts. The first part starts in the general field of NLP, presenting an overview and highlighting recent developments through selected language models relevant to the work of this thesis. The aim is to emphasize context for the reader to better tune into the work in the forthcoming chapters. The second and third part zooms in on financial NLP, covering relevant literature and challenges that have motivated the work presented in this thesis.



Timeline overview of language models considered in background 2.2.1.

2.2.1 *Representation Learning in NLP*

Pretrained word embeddings

WORD2VEC The release of WORD2VEC in 2013 by Mikolov et al. [144] shifted the focus of the NLP community and popularized pretrained continuous word embeddings.

Learning a vectorized representation of words is required for many machine learning tasks. A word embedding is a vectorized representation of a word as a real-valued vector that encodes the meaning of the word mapped to a vector space [106]. Pretrained word embeddings offer substantial improvements over embeddings needed to be re-learned from scratch each time and discrete text representations, such as Bag-of-Words (BoW) [78, 143, 144]. WORD2VEC learns continuous word representations through unsupervised learning from a large unlabelled corpus. The main contribution of WORD2VEC is that the embeddings are in a much lower-dimensional space than sparse embeddings (e.g., BoW), and these dense embeddings are better at capturing semantic relations between words. WORD2VEC can provide estimates of a word's relations with other words based on its occurrence in a large corpus containing sequences of words. This is an influential advance since words with semantic similarity will have similar vectors.

WORD2VEC uses two model architectures to learn the underlying word representation: continuous Skip-Gram and continuous Bag-of-Words (CBOW) [143]. The objective of the Skip-Gram model is to predict the surrounding context words from the target word. The context is the set of nearby words to the target word controlled by the window size. Contrary to the operation of the Skip-Gram model, CBOW's objective is to predict the target word based on the surrounding context words. This makes WORD2VEC able to learn the probability of seeing a context given a word, or a word given the context [143, 144].

The effectiveness of grouping together vectors of similar words drives the advantages that WORD2VEC has compared to the shortcomings of previous approaches [18, 78, 85]. Sparse embeddings such as BOW [85] lack useful semantic properties, the ability to induce proper similarity between words and have the same dimensionality as the number of distinct features in the vocabulary [18, 78]. However, one shortcoming of WORD2VEC (and BOW approaches) is that these methods cannot produce meaningful vectors for words that did not appear during training, referred to as out-of-vocabulary or unknown words. Consequently, this shortcoming poses a challenge if a new, unknown word appears in the test corpus.

FASTTEXT In 2017, Bojanowski et al. [22, 81, 104] presented FASTTEXT, which marked an important advancement in NLP as it allows for computing meaningful representations of words that did not appear during training. FASTTEXT has outperformed baselines not considering subword information (character n-grams) [22].

FASTTEXT alleviates the problem of words being out-of-vocabulary by introducing an extension to the work presented by Mikolov et al. [143, 144]. Bojanowski et al. [22] propose an extension to the Skip-Gram model [143, 144] that also accounts for subword information as an alternative to using distinct vectors for representing each word as a whole. With embeddings using subword information instead of word-level embeddings, each word is represented as a bag of character n-grams[7], where a vector representation is associated with each n-gram. This defines a word's embeddings as the sum of the embeddings of its character n-gram [22] and makes it possible to obtain a word vector for unseen words as long as it can be split into character n-grams observed during training. Furthermore, a limitation of prior work that FASTTEXT improves is better modeling of internal structures of words, enabling better processing of languages with large vocabularies, rich morphology, and rare infrequent words.

Due to an increased interest in multilingual NLP, Grave et al. [81] released word vectors in 157 languages, underlining the importance of making FASTTEXT applicable to tasks in multiple languages. This contribution stresses the significance of resources in the NLP community, not only in English, but in several languages.

Models such as WORD2VEC and FASTTEXT produce static embeddings, meaning that each word in the vocabulary is assigned a fixed embedding. In contrast, models such as ELMO and BERT learn dynamic contextual embeddings, where the word representation is de-

[7] Bojanowski et al. [22] gives an example of n=3, representing the word <WHERE> as the set of character n-grams: <WH, WHE, HER, ERE, RE>, including special boundary symbols to distinguish the start and end of a sequence. The tokenization of <WHERE> creates the shown character tri-grams.

pendent on the context, resulting in a different vector in different contexts [22, 106, 164].

Contextual embeddings

ELMO: EMBEDDINGS FROM LANGUAGE MODELS Introduced by Peters et al. in 2018, ELMO [163, 164] is considered an important advancement in NLP as it brought contextualization into focus and showed impressive results on standard benchmarks. Through *deep contextualized* word representations, ELMO aims to address the complex characteristics of word use such as syntax and semantics and how they vary across contexts [164]. ELMO is a bi-directional language model composed of a long-short-term memory network (LSTM) [88]. It is a variant of recurrent neural networks capable of better learning long-range dependencies. It is beneficial for sequence processing tasks as found in NLP since natural language sequences often encode such dependencies, and the processing of a word often depends on the words at previous positions. As described by Peters et al. [164], each token can be assigned an embedding being a function of the entire input sentence, which stands in contrast to previous static word embeddings [22, 81, 144]. Previous static word embeddings cannot express the context in which a word is used. In contrast to FASTTEXT and WORD2VEC, ELMO's embeddings are context sensitive resulting in different representations of a word given different contexts. That is, the embeddings for "*transaction*" may be different for "*financial transaction*", "*bank transaction*" and "*electronic transaction*".

Pretrained language models

TRANSFORMERS In 2017, Vaswani et al. [224] proposed a new network architecture based on the so-called "attention" mechanism, the *Transformer*, entirely omitting the use of recurrence and convolutions, which represented the state-of-the-art at the time for sequence modeling, e.g., language modeling and many NLP tasks. Vaswani et al. [224] showed that the transformer achieved superior performance on two machine translation tasks while being more parallelizable and thereby faster to train compared to recurrent models. The recurrent model processes text sequentially, that is, one token at each step, whereas the transformer processes all tokens simultaneously. In a recurrent neural network, there is the problem of loss of information for long sequences, and the ability to learn long-range dependencies between token positions decreases with distance [106]. The transformer's attention mechanism allows for modeling dependencies without concern for their distance in the input sequence. When processing each item in a sequence, the attention mechanism allows the model to access the inputs at any position along a sequence. The transformer can simultaneously calculate the attention weights between every token and create embeddings for each token in context. The attention layers can draw from all states and provide relevant information about distant tokens [224]. The transformer model shifted the NLP community's focus away from the recurrent neural network architectures that had previously being regarded as the state-of-the-art to transformer-based architectures [46, 47, 59]. The transformer model addressed previous problems and created a new direction for language models. Shortly after, pretrained transformer-based models, such as BERT [59] and XLM-R [46, 47] started to fuel a new shift in the field of NLP.

BERT: BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMERS Introduced in 2018 by Devlin et al. [59], BERT showed that a single model could achieve state-of-the-art results on eleven of the most common NLP tasks. It presented a new paradigm for language modeling and made a significant advancement in NLP.

The architecture is a multi-layer bidirectional transformer encoder almost following the original model presented by Vaswani et al. [224]. BERT is pretrained with the masked language modeling (MLM) objective, i.e., cross-entropy loss on predicting the actual tokens for randomly masked tokens. MLM is a pretraining objective that works by randomly masking a low percentage of the tokens from the input sequence with the objective to predict the actual original token being masked, given the remaining surrounding tokens in the input sequence. In addition to MLM, BERT also uses next sentence prediction (NSP) as a training objective. NSP is a binary classification loss for predicting the next sentence that follows after the current sentence, i.e., whether a segment ISNEXT or NOTNEXT. Devlin et al. [59] used the NSP training objective to pursue a better understanding of the relationship between sentences, as it could improve essential aspects for several downstream tasks.

The pretrained BERT model is trained on vast amounts of data in an unsupervised way, allowing the model to learn a rich representation of language that can be reused in a variety of downstream tasks [59]. It follows the paradigm of *pretrain-then-fine-tune*. For fine-tuning BERT, for example, for text classification, a classification layer is added to the pretrained model, and all parameters can be fine-tuned on the downstream task. Alternative to fine-tuning, word-level or sentence-level features can be extracted from the pretrained model and used downstream in a feature-based approach similar to FASTTEXT and WORD2VEC [59]. Similar to FASTTEXT, BERT can produce meaningful vectors for unseen words as it embeds subwords rather than words, and BERT also accounts for a word's context like ELMO does.

Numerous studies have investigated BERT [169, 193]. Liu et al. [124] investigated ways of optimizing BERT's pretraining approach and found that they could further improve BERT. The new version, RoBERTa, obtained new state-of-the-art results on multiple benchmarks and outperformed the original BERT. Liu et al. [124] pretrained over a 10x larger corpus with fewer episodes over the data and used a nearly 8x bigger batch size. They also tested the necessity of NSP and found slightly improved performance when they removed the NSP training objective. This shows that further improvement can be obtained from training on more data for longer time [124]. These insights inspired Conneau et al. [46] to introduce XLM-RoBERTa, referred to as XLM-R in the next segment.

Pretrained multilingual language models

MBERT: MULTILINGUAL BERT After the release of BERT, Devlin et al. [59] also released a multilingual BERT pretrained on text in more than 100 languages. Rather than pretraining on a corpus of English texts, Devlin et al. [59] pretrained on a corpus of texts in multiple languages using the same BERT model, but with multilingual data without explicit cross-lingual supervision. It attracted much attention and is extensively used in the literature for multilingual NLP, given its effectiveness in cross-lingual and zero-shot cross-lingual transfer. Cross-lingual transfer refers to transferring knowledge learned on data in a source language to data in a target languages [46, 91, 169, 192].

XLM-R: CROSS-LINGUAL LANGUAGE MODELS Introduced in 2019 [46, 47], XLM-R is considered an important advancement in the field of NLP and it intensified the focus on multilingual NLP, demonstrating the possibility of multilingual modeling without sacrificing per-language performance. The study on XLM-R showed that large-scale pretraining of multilingual language models can significantly improve performance for a broad range of cross-lingual transfer tasks and be competitive with monolingual models. Besides presenting state-of-the-art results on benchmarks, Conneau et al. [46] showed that multilingual models can outperform their monolingual counterparts, i.e., XLM-R outperformed BERT on monolingual benchmarks and was found competitive with other monolingual models [46]. This moved the focus of the NLP community toward general-purpose multilingual representations.

XLM-R is a transformer model trained with the MLM objective, similar to BERT [59]. The XLM-R model closely follows the approach of XLM [47] and RoBERTa [124], therefore dubbed XLM-RoBERTa. The study by Conneau et al. [46] further points out some details of multilingual modeling, such as the fact that multilingual language models are limited by their capacity. Given a fixed-sized model, as the number of languages increases, the per-language performance decreases. Conneau et al. [46] defines this as the *curse of multilinguality*. This phenomenon indicates that larger models may be needed to obtain comparable performance to monolingual counterparts. The XLM-R study also confirmed the finding by Liu et al. [124] that showed the benefit of training language models longer and with a larger pretraining corpus.

Massively pretrained language models

GPT-3 AND NEWER MODELS Research in the field of NLP continues to build on the above findings, presenting ever more powerful machinery [28, 43, 44, 207].

Model	# parameters
LAMBDA [220]	137B
GPT-3 [28]	175B
GOPHER [177]	280B
MT-NLG [207]	530B
PALM [44]	540B
XLM-R [46]	270-550M
BERT [59]	110-340M

Table 2: Sizes of state-of-the-art NLP models.

Models are trained increasingly longer over more data with increased model size, demonstrating that scaling up language models greatly improves performance [28, 43, 124].

Table 2 shows the trend of sizes of state-of-the-art NLP models. These models continue scaling up in the number of parameters with sizes significantly surpassing BERT and XLM-R. For instance, GPT-3 contains 175B parameters which is a considerable scale-up compared to BERT with 110-340M parameters and XLM-R with 270-550M parameters.

Despite achieving impressive performance, massively large-scale language models are becoming more complex and often not available to the public, e.g. behind pay-to-access APIs. In addition, they are too costly to share and serve for most institutions. These were essential motivating factors for the work conducted during this thesis.

2.2.2 *Models in Financial NLP*

While the previous section emphasized research milestones and context in the general field of NLP, this section concentrates on representation learning and models in the domain of financial NLP. Within the field, there is a focus on learning general financial language representations from domain-specific data, in addition to advancing the state of downstream tasks and challenges within this domain.

REPRESENTATION LEARNING IN FINANCIAL NLP Financial NLP, like other domains, benefits from general-purpose language models and general advancements in the field of NLP. Although recent massively pretrained language models achieve strong performances across many tasks and have been shown to be good general-purpose models, it is still beneficial to specifically focus on in-domain adaptation [83]. A number of studies confirm that in-domain and domain adaptive pretraining is helpful for several domains [35, 83, 120], including the financial domain [8, 54, 101, 126, 239]. Domain-adaptive pretraining is defined as continuing to pretrain a language model to fit a specific domain [83, 84], as an alternative to pretraining a language model from scratch. Continued pretraining (CPT) may be a preferred domain adaptive strategy compared to training from scratch when it is not possible to find vast amounts of domain-specific unlabeled texts to pretrain from scratch. Also, even if vast amounts of domain-specific data were available, CPT might still be the preferred option because it requires less computational resources. CPT benefits from the general-domain as the general-domain may have an abundance of available data that can be used to learn a good base model. The aim of CPT of mono- or multilingual base models is to adapt the general-domain model to the idiosyncrasies of the specific domain, while preventing the base model from forgetting, e.g., how to represent multiple languages [8, 83, 90, 101]. A growing body of literature concentrates on representation learning using transformer-based models, where the BERT model is trained using financial corpora. The term FinBERTs is used here to refer to the models that result from this line of work [8, 54, 126, 239]. These different versions of FinBERTs are either pretrained from scratch or CPT. These monolingual models that are trained using corpora of financial text in English demonstrate the benefit of domain-specific adaptation and confirm the original work by Gururangan et al. [83].

Besides FinBERTs, a number of studies within the community have concentrated on particular downstream tasks or domain-specific challenges, for instance, number-aware languages models [122], models for financial numeric entity recognition [131], including datasets focusing on specific interests of financial NLP, such as AnalystToneDataset [92], FinTextSen [50] and Financial Phrase Bank [138], and others as shown in Table 1. Moreover, there is a pronounced research interest in analyzing financial news and market dynamics [39, 56, 63, 233], including textual transcripts [176] from institutions and policy makers [4, 29, 141], volatility and market return [60, 77, 196, 217], risk and uncertainty estimation [3, 125, 128] as well as forecasting firm performance [218] and bankruptcy [135], inter-alia.

As argued in this section, many of these lines of research within financial NLP could benefit from improved financial language models. Although it is difficult to grasp all developments within the large communities of financial NLP and the general field of NLP, there are a number of differences in the literature of the two fields. The literature on financial NLP does not contain much work on different transformer-based models or alternative models, as covered in 2.2.1. While the general field of NLP compares and contrasts different models on downstream tasks and benchmarks, the financial domain does not have similar counterparts available for comparison and evaluation [8, 54, 101, 126, 239]. Furthermore, the literature on financial NLP primarily uses monolingual approaches in English and does not contain much work on multilingual models. Although English has become *lingua franca* in the business domain, the financial domain is inherently multilingual by nature as text and information occur in multiple languages [99, 101].

MODELLING FINANCIAL TRANSACTIONS One line of work within financial NLP concerns processing text from financial transactions. This financial task is different from typical NLP tasks in several respects. Transactions contain several features, as presented in a tabular form in Table 3, where text fields are a feature subset. The accounting task focuses on accurately mapping transactions to their accounts. The emphasis in the area is to extract predictive features from the transaction text to support classification in combination with the other features. Transaction text is characterized by free-text fields either machine-generated or filled-in by humans. The text consists of short incomplete sentences containing fragments without much context and sometimes domain-specific abbreviations and improvised words, often including references to, e.g., invoices [19, 70, 102, 225].

Date	Transaction Text	Amount	Account Code
07/04/22	PS8877 internet company	-289	270 (Utilities)
12/05/22	Sale*StoreName 3834 X products	231	100 (Sales)
04/06/22	fruits from StoreName (Employee No.)	-93	223 (Other expense)
18/06/22	Invoice-3231 PlumberName ltd.	-385	273 (Maintenance)
25/06/22	DK5545 grocery store copenhagen	-56	220 (Other expense)
27/06/22	Invoice-2321 3435-1231231	-543	? (Needs review)

Table 3: Fabricated examples of financial transactions that highlight some of the text variation that a system needs to process.

While these text fields are scarce in their information, they often contain enough information to infer a mapping to a class [19, 20, 206, 225]. Research concentrates on how best to exploit the unstructured information in these text fields [70, 71, 225], and in the literature there has long been interest in processing text for accounting tasks and addressing domain-specific text in this area. In 1987, Mui and McCarthy [148] stated an early interest in applying NLP techniques for analyzing and automating accounting tasks [148]. In 1992, O'Leary and Kandelin [157] published research on domain dependent accounting language processing system [157]. More recent attention on text processing for financial transactions has mainly focused on BoW approaches and simple pattern matching [19, 70, 71, 225]. There is a relatively small body of literature concerned with techniques to enrich transactions with external information [20, 70, 71]. Most of the highlighted studies utilized text features as input to machine learning models, such as SVM, logistic regression, or neural networks. Also, exact match, BoW-based methods, and other count-based representations were considered for pattern matching and hand-crafted rules [19, 20, 85, 102, 112, 148, 206].

2.2.3 *General Considerations in the Thesis*

The work presented in this thesis contributes to improving the modeling of multilingual text for downstream tasks within the financial domain. Chapter 5 considers multilingual domain-adaptive pretraining, i.e., the continued unsupervised pretraining of a multilingual language model on domain-specific text in multiple languages.

The reviewed literature is mainly concentrated on monolingual approaches, such as FinBERTs. Although trained on domain-specific text, these approaches are mainly English-centric, given the vast availability of English resources. The importance of multilingual language representations was highlighted in 2.2.1, inspiring the research in this thesis to narrow the gap in the literature by advancing multilingual NLP in the financial domain. In contrast to current developments in the field of NLP, we concentrate on CPT approaches and ways of combining multilingual pretraining data accessible to most institutions and practitioners. Altogether, the research aim is to contribute to a deeper understanding of multilingual domain-specific models in the field of financial NLP. While this thesis focuses on the financial domain, methods and findings presented here could, to some extent, also be applied to improving multilingual NLP in other specific domains, exemplified by the biomedical experiments shown in chapter 5.

Chapter 4 focuses on addressing the challenges introduced in processing transaction free-text fields. For this, we work on a transaction representation with the end classification task of mapping transactions to accounts. The nature of the unstructured transaction text, as covered in 2.2.2, contains a combination of words that often enables a transaction mapping to a specific account. In the text, words may often co-occur and co-participate in a mapping to an account, for example, "PS" (automatic payment service) and "phone" often makes mapping to an account containing corporate phone expenses trivial. Besides word relatedness, the ability to capture similarities and associations among words is important, as similar transactions should have similar representations [78]. For example, "fuel", "gasoline" and "diesel oil" most likely all belong on the account for "Transportation". In terms of generalization, new unseen words should be related to neighboring words, as a more suitable neighborhood measure between the words may induce helpful associations between transaction instances [102].

The currently used approaches revealed certain limitations [19, 20, 70, 71, 102, 225]. For instance, a challenge appears when using count-based lexical representations, such as BoW or a rule-based approach[8], as they do not generalize to words not in the training corpus or to undefined cases and do not exploit similarities. Another prominent challenge is the curse of dimensionality [18] when extracting features from text to be combined with other features, as the dimensionality grows with the number of words in the vocabulary. This makes short, dense embeddings a promising alternative compared to sparse embeddings with dimensionality as their vocabulary.

[8] As we consider in chapter 4, a rule-based approach refers to simple pattern matching methods, such as human-made rules for text matching keywords. For example, if "phone" appears in text, then map the transaction to account for "phone expenses" [19, 20, 102].

Inspired by the above observations, this study seeks to identify more suitable methods for handling transaction text. Firstly, this study focuses on an embedding that provides a more refined distance measure between the words and across transactions so that similar transactions will have similar vectors. This study investigates whether better generalization is fostered by inducing a more helpful neighborhood similarity between transactions. Secondly, this study seeks to bring forward the challenge of constructing meaningful vectors for unseen words outside the training corpus. Lastly, although not emphasized in the literature, it is useful if a suitable approach could be able to extend to multiple languages since financial transactions are processed globally. Chapter 4 details the present study's investigation on processing transactions for the aforementioned accounting task of mapping transactions to accounts.

2.3 EXPLAINABILITY IN NLP

This section will provide supplementary information on the methods used in this thesis to explain machine learning models. Because the use of the selected explainability methods has become common, and this thesis uses them in their original form, the background does not intend to be a detailed and comprehensive presentation – Søgaard [210] can be referred to for an extensive review of explainability methods used in NLP. In the recent pertinent literature, interpretability refers to the ability to understand what a model has learned, e.g., inspecting the weights of linear regression [146], while explainability refers to the ability to provide an explanation of *"why a model produces a certain prediction for a certain instance"* [12].

Generally, explainability of machine learning models is a broad domain that spans different areas of research. This dissertation focuses on feature attribution-based methods and evaluating explanations with human rationale annotations.

Large-scale pretrained language models dominate the field of NLP and are widely used in both research and industry due to their superior performance compared to simpler models such as k-nearest neighbors or decision trees. These larger models are becoming increasingly complex and this trend seems to continue [28, 43, 46, 59]. The increased model complexity makes it more complicated to understand how a model has arrived at its prediction, which consequently creates a demand for more explainability. The desire for superior performance and simultaneously high explainability creates the *trade-off between performance and explainability*, since it is difficult to obtain both properties in a single model [30, 133, 187]. The aim of explainability is to address this need by providing an explanation behind a particular prediction (e.g., from a complex model). For some researchers and practitioners, particularly non-technical users of such systems, it is imperative to understand the reasons behind the predictions made by an NLP system [30, 123, 133, 187]. One example where explainability is relevant is in the medical domain [132, 211, 222], where the medical staff need to assess the reasons behind the predictions. Another example is the constraints imposed by the regulations around machine learning applications [31, 51, 165].

Within explainable natural language processing, a wide range of explainability methods and approaches to explainability evaluations exists [210]. Current explainability research in the NLP community and across other machine learning domains is very active, and it develops together with adjacent topics such as evaluations [55], definitions [96, 123], bias and fairness [116, 142, 154], *inter-alia*.

This thesis considers SHAP[9] [133] and LIME[10] [187] methods for the study in chapter 7. Besides these two methods, other types of popular explainability methods are used in the machine learning community,

[9] SHapley Additive exPlanations (SHAP) by Lundberg and Lee [133].

[10] Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro, Singh, and Guestrin [187].

such as DEEP TAYLOR DECOMPOSITION [147], LAYER-WISE RELEVANCE PROPAGATION [14], and many others [30, 146, 210].

2.3.1 *Post-Hoc Attribution-Based Methods*

LIME and SHAP are considered core contributions to the field of explainability [30, 146, 210]. They share many similar properties and belong to the same category of *post-hoc model-agnostic* feature attribution methods that assign an importance score to each feature of an input based on a feature's contribution to the model's prediction. The highlighted feature attribution methods are *local methods* that aim to explain the individual prediction, while global methods seek to explain the entire model behavior. *Model-agnostic* means that the explainability method does not rely on assumptions about a specific model but instead separates the model from the explanation. These methods treat the complex model as a black box that can be probed, e.g., for probabilities of a prediction. A *post-hoc* method refers to a method applied after model training. That is, the method is applied to inspect the prediction made by the trained model based on the predicted example.

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME) In 2016, Ribeiro, Singh, and Guestrin [187] presented LIME as a model-agnostic technique to explain a model's prediction.

LIME's goal is to approximate the complex model's prediction through an interpretable surrogate model, e.g., by fitting a linear model on perturbations of the example to obtain a local approximation of the model's decision boundary around the predicted example. Thereby, LIME approximately explains a complex model by using a simpler and more interpretable model. In the case of text, LIME probes the model by providing variations of the text instance through perturbations obtained by randomly removing words or characters [146, 172, 210].

Given the original example x and the original complex model f, we want an explanation of $f(x)$. Ribeiro, Singh, and Guestrin [187] define the explanation $\varepsilon(x)$ obtained by minimizing the loss by using a linear surrogate model g with interpretable constraints as

$$\varepsilon(x) = \mathrm{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \ .$$

Perturbed samples z are sampled based on x, i.e., binary vector representation of features indicating presence or absence of a word. An explainable model is defined as g from a class of possible interpretable models G. The loss $L(f, g, \pi_x)$ defines how successful g is in approximating f for the locality defined by π_x, which is used as the proximity measure between the perturbed z and original example x. LIME minimizes the loss to obtain an approximation to the original model. The measure $\Omega(g)$ is the complexity of the model and must be picked, for example, in a text setting, by setting a limit on the number of words [146, 187]. The choice should be low enough to ensure a good explanation without placing too high a demand on computational resources. $\varepsilon(x)$ represents the explanation of x under the constraint of being a good local approximation of $f(x)$.

LIME has been subject to further research [30, 146, 187, 210], with work in tailoring a model-specific fit to convolutional, graph and recurrent neural networks [93, 172], redefined as a quadratic function [25], and different sampling variations [114, 200], including substring-based instead of word-based sampling for text [172]. Variants of LIME have been used in many domains including the audio [87], text [210] and medical domains [211, 222].

SHAPLEY ADDITIVE EXPLANATIONS (SHAP) In 2017, Lundberg and Lee [133] presented a method for interpreting machine learning models, SHAP, that builds on Shapley values. It spurred numerous applications within the machine learning community [190]. Shapley values [201] is a method with a theoretical background in coalitional game theory. The concept hinges on distributing the *payout* among the *players* in a coalition based on their contribution. Lundberg and Lee [133] re-frames the game theoretical concept such that each feature of an example is considered to be a *player* and the *payout* is the sum of feature importance from each feature to the predicted example. The *game* is to reproduce the outcome of the model with the *payout* as the marginal contribution of a feature. The marginal contribution per feature is quantified through the averaged effect of all possible feature combinations and how they contribute to the prediction [146]. For text examples, SHAP works by removing words or characters and produces explanations in terms of words or characters.

Let x be the input sequence of words and f the original complex model. Lundberg and Lee [133] define z' as simplified inputs being the binary vector representations of all possible combinations, with words either present (1) or absent (0) in a particular combination. It is helpful to think of z's as different combinations of included or excluded words from the input sequence x. Given the explanation model g, that is a linear function of binary variables, Lundberg and Lee [133] present the explanation as:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \ ,$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, ϕ_0 is the null output, and ϕ_i is the feature attribution for feature i, i.e., Sharpley values. This defines the marginal contribution of each word feature in a sentence by summing the effects towards the model's predicted outcome. The goal is to approximate the original model $f(x)$ as closely as possible with the explanation model $g(z')$, but the exact computation of ϕ_i is in most cases infeasible. The challenge posed by calculating the marginal contribution of each feature combination through Shapley values is that the more features a model relies on, the higher the number of coalitions becomes, making it not scalable to current developments in machine learning [146].

However, approximative methods are presented in the SHAP implementation [133]. Among the different methods used are KernelSHAP, which is model-agnostic and uses a linear model for local approximations, LinearSHAP which is used for linear models approximated directly from the model's weight coefficients, and TreeSHAP that measures local feature interaction effects for trees [132]. In extension, PartitionSHAP, PermutationSHAP, among others, are also used to improve computational efficiency and reduce the number of coalitions used to calculate the Shapley values. For a more elaborate presentation of Shapely values, we recommend Shapley [201], and for SHAP, we refer to the original work by Lundberg and Lee [133].

2.3.2 *Evaluation using Human Annotated Rationales*

Explainability research devotes considerable attention to defining ways of evaluating explanations with the objective of benchmarking explainability methods. One direction of research within evaluation in NLP is using *human rationales*. Human rationale annotations can be used to evaluate the extent to which a model-generated rationale (explanation) aligns with a human rationale as a metric for assessing different explainability methods across various tasks, datasets, and desired criteria [55, 133, 187, 210].

Annotation guidelines for human rationales are described by Zaidan, Eisner, and Piatko [242], which details how humans are asked to highlight the parts of the text that are considered important for predicting the individual instance. Giving a positive movie review as an example, which factors of the review determine that the review is positive? Human rationales can be used as ground truth in evaluating how a model-generated rationale aligns with human rationale for an individual prediction. That is, marking the snippet of text that supports the outcome; and whether the provided explanation is in agreement with the human rationale annotation.

Research and resources include the benchmark dataset ERASER [55] and task-specific evaluation datasets containing human rationales for different tasks, such as natural language inference [32], sentiment analysis [241], hate speech detection [140], and fact-checking [221]. While this line of research primarily considers *plausibility*, referring to how convincing the explanations are to humans [96, 123], others use rationales differently, such as for partly supervising model training by learning from these rationales [214, 242, 245] and for data augmentation [231].

Chapter 7 concentrates on building a multilingual corpus of human rationale annotations in Danish, Italian, and English, for the task of sentiment analysis that can be used to benchmark multilingual language models and explainability methods.

3 INDUSTRIAL COLLABORATION

The industrial research collaboration with PwC focused on advancing their internal R&D agenda through research. While the previous section discussed the research aspect of my PhD work, this section concentrates on the industrial collaboration, which was an important aspect of this PhD study that involved identifying problems and designing projects to address these problems. This was followed by collection of data, implementation and hand-over of solutions to PwC, which took up a large part of my time during the PhD. This section highlights the contributions of my PhD in relation to the industrial research work and the solutions implemented at PwC.

3.1 PROJECT TIMELINE

The industrial collaboration was structured around two phases with each phase being a topic. The two phases are:

PHASE 1 (Month 0 – 18) Classification of Financial Transactions

PHASE 2 (Month 18 – 36) Multilingual Financial NLP

Common for both phases is the initial planning in which challenges were identified and a project was designed with a business need in mind. This was followed by preliminary research and data collection before the research project started. As the research was concluded, it was handed over to the industrial partner for implementation and integration. The remaining parts of this section will concentrate on the industrial aspect of each topic and since it is similar to the content of the papers that chapters 4, 5, 6 and 7 are based on, it may contain passages from the papers without quotation.

3.2 CLASSIFICATION OF FINANCIAL TRANSACTIONS

This segment covers some of the industrial aspects behind chapter 4 containing the research work in [102]. The industrial partner is often commissioned for accounting and bookkeeping services. One such activity is accounting financial transactions, i.e., mapping transactions to accounts and summarizing a company's transactions. Although companies have become highly digitized, accounting and bookkeeping still require much done by human manual effort. Repetitive and standardized activities characterize the process of mapping transactions to accounts, and this is often partly automated by rule-based systems. Experts need to maintain these rule-based systems, and these systems often leave a considerable portion for post-correction. Maintaining high quality is resource-demanding in terms of time and costs.

Therefore, accounting firms would benefit from an automatic system that learns to classify transactions across companies and sectors, and that is also able to generalize to new companies for which little or even no historical transaction data may yet exist [102].

PREPARATION AND DATA COLLECTION Having identified the above research topic and business problem, we started working on data collection and unifying the chart of accounts[1]. This presented us with several initial challenges. Data were not easily accessible and special extraction mechanisms needed to be defined for collecting the data. After the collection of data, we found that the structure and format of the data were tailored to the individual companies. Thus, it required much preprocessing, such as reformatting and sorting out invalid data. Similar to the data, the chart of accounts was tailored to the individual companies. For the unification, we needed to recruit the accounting team to define a standardized chart and subsequently establish the mapping between the individual to the unified chart. This demanded a lot of resources and time, especially since data collection was time consuming, and unifying the charts of accounts was an iterative process requiring much involvement from the accounting team.

INITIAL STEPS AND FEEDBACK FROM EXPERTS We first aimed to address the need for an input and output space by creating a transaction representation and a representation of the chart of accounts. We increased the number of features through feature engineering from three to twenty-eight, for example, extracting text features, payment type, the distance between the two entities, etc. The methods used helped create a transaction representation that improves the system compared to the standard features. The unification of output spaces and engineered features enabled us to learn across companies utilizing transfer learning. The unified chart of accounts is sufficiently general, and the availability of pretrained word embeddings in other languages makes it transferable to different types of companies and languages. While the base system targeted at individual companies only assumed that data from a single company were available for the system, the extended system enables classification of financial transactions across companies and corporate sectors. This allows the system to better tackle the *cold-start problem*, also making the system applicable to companies for which no or very little historical data are available. For both scenarios, the possibility of updating the system by retraining the models is considered an advantage compared to rule-based systems.

Eight experts[2] in accounting and bookkeeping at PwC gave feedback on the work. The expert feedback evaluated these investigations

[1] Accounts record a company's transactions that together compose a hierarchy defined by the *chart of accounts*. The chart of accounts is not always standardized across companies but often tailored to the specific company. Therefore, it hinders classifying financial transactions to accounts across companies. [102]

[2] Two senior associates, three senior managers, two directors and one partner.

to provide solutions to their challenges for advancing toward more automated systems. This system is also able to improve the quality compared to the rule-based setup. The expert feedback showed that the solutions developed in this PhD study match the defined business need and address both the scenario of individual companies and groups of companies. It was appreciated that the proposed methods are not too resource intensive. The industrial partner estimated these solutions could save approximately 30-50 percent in costs and time compared to current rule-based systems [102].

PROTOTYPE Because of the high accuracies, two prototypes were developed in extension to the research project: one for an internal pilot test and another for an external test case. The internal prototype system was developed as a web application and tested using PwC's in-house historical data previously used in the accounting process. The prototype demonstrated how transitioning to a machine learning-based approach improves the work process compared to the existing rule-based systems. The research work also resulted in a prototype that was externally tested. It was tested on data from a larger company together with the industrial partner. In addition to developing the prototypes with PwC, we also addressed the system design, deployment of the machine learning-based systems and how it should interact with accounting and bookkeeping clerks - for instance, whether it should be fully automated, semi-automated, or a top-5 recommendation to the bookkeeping clerks.

FURTHER R&D AT PWC The research work described above establishes the basis for further research and development at PwC. The different datasets for model development, targeting either *companies individually* or *across companies*, enable future work on designing new models. The internal benchmark of collected transactions, including various evaluation scenarios, allows for comparing models across different scenarios on a standardized testbed. As to the unified chart of accounts mapping for aligning companies' output spaces, we also identified a way to transform them using the existing system in a more elegant integration. At PwC, this line of work shows promise as a solution to the industrial problem, a match with the defined business need, and suggests future directions of research to this industry challenge.

3.3 MULTILINGUAL FINANCIAL NLP

Processing large amounts of textual information is important for accounting, tax and other financial services, where text is processed in multiple languages for tasks such as invoice and transaction processing [102], processing texts for auditing [129], classifying and retrieving information from documents for legal, tax and investment analysis [54, 129, 217, 218]. All these tasks contain both a very specific use of language and terminology that is not in the general-domain and common language used for training general-purpose models [46, 47, 59, 101].

PwC or any similar firm cannot assume exclusively English text or any other monolingual setting, e.g., the national language, but needs to process texts in multiple languages, even for local assignments, e.g., minimum English and the national language. This makes the field an inherently multilingual environment, where deployed systems should be prepared to process domain-specific documents in several languages. This circumstance has created a growing interest and a need for multilingual, domain-specific language models.

As this line of work has progressed, business needs have arisen that require solutions to challenges generally met in the financial domain when working with multilingual natural language processing. Together with the industrial partner, this PhD project developed to address these needs, focussing on the following: 1) DATA RESOURCES, 2) MODEL DEVELOPMENT and 3) EVALUATION. Step 1 was to establish the prerequisite data resources for model development and assessment of financial NLP models. Step 2 was to investigate how to train and evaluate domain-specific multilingual models. Step 3 was to evaluate the output of deployed models and explain these to potential users of the multilingual NLP system.

DATA RESOURCES AND DOWNSTREAM TASKS Establishing appropriate data resources for training and testing models is important for work on multilingual NLP in the financial domain. A common shortcoming in this area is the lack of available labeled and unlabelled domain-specific data in multiple languages. To address this issue, it was necessary to collect a vast amount of domain data. Subsequently in this line of the PhD study, test cases had to be defined and evaluation tasks produced in multiple languages. Building this foundation to address the research topic and business need entailed a time-consuming and labor-intensive process with regard to data collection and annotation. We invested time in developing the following resources:

FINMULTICORPUS (CH. 5) is a pretraining corpus of financial texts in 14 languages. We initiated a larger collection of articles, books and other texts to build a large unlabelled pretraining corpus.

DANFINNEWS (CH. 5) is inspired by the popular FINANCIALPHRASEBANK [138]. We produced a Danish version for evaluating models on a classification task with Danish domain-specific text.

MULTILINGUAL FINANCIAL BENCHMARK (CH. 6) is a multilingual benchmark dataset for evaluation of multilingual domain-adapted models. It is a real-world financial dataset covering 15 languages across different writing systems and language families.

SST WITH HUMAN RATIONALES (CH. 7) is a multilingual dataset with human annotated rationales. We created a trilingual parallel corpus of human rationale annotations in Danish, Italian, and English, for the task of sentiment analysis using Stanford Sentiment Treebank (SST) [209].

The creation of these resources required a large amount of time. The quality was a high priority, which required sometimes redoing annotations and definitions for quality assurance and consistency. Besides the annotation, significant planning went into obtaining permissions, extracting data and defining annotation schemes. Particular effort was devoted to organizing volunteers for annotation and review, including planning work, guidelines and the like. This line of work also provided knowledge and experience to the industrial partner for future projects on creating high-quality datasets necessary for developing a machine learning system.

TRAINING DOMAIN-SPECIFIC MODELS In chapter 5, we suggest a method for producing domain-adapted multilingual models, including strategies for the composition of unlabelled data for continued pretraining datasets. Also, considerations around limitation and feasibility are taken into account since data availability and compute cost must be evaluated for all industrial projects. As promising industrial projects may be rejected because time and compute resources are too high, we considered how to train a language model for a new domain while working with resources on a budget. We investigated different strategies for composing a pretraining dataset in a situation where it may not be possible to obtain sufficient samples in different languages. The industrial partner was pleased with the work presented in chapter 5 and evaluated that it addresses their technical needs and challenges while assuming a reasonable availability of resources. In particular, the strategies for composing pretraining datasets were considered beneficial for many situations, and the fact that a single model eases deployment makes it more useful across different projects.

EVALUATION OF NLP SYSTEMS Chapter 6 considers the evaluation of NLP systems in terms of performance. We created a financial multilingual benchmark dataset for evaluating domain-adapted models across 15 languages of different writing systems and language families. The evaluation dataset can serve as a part of the testbed at PwC for assessing which NLP systems and setup should proceed to deployment.

When an NLP system is deployed, interactions with these systems may sometimes require humans to inspect predictions, or in some cases require them to do so constantly, if a system is built solely to assist humans in their work. For instance, if a document is flagged as sensitive, a data protection officer may want to review the reasons behind the classification before starting to process the document and make a decision about the case. For such purposes, explainability methods can be useful and provide a supportive functionality for the users of the NLP system. Testing NLP systems goes beyond accuracy [188], and insight into how these explainability methods evaluate with human users are important considerations. The work presented in chapter 7 was initiated to obtain knowledge of how selected explainability methods evaluate with human rationale and insights into explainability in multilingual environments.

INDUSTRIAL CONTRIBUTION AND FUTURE R&D AT PWC The objective of this line of work was to mature the foundation at PwC for developing machine learning applications for financial NLP with respect to research, development, and deployment of NLP systems. Through the research carried out in the real-world setting of the industrial partner, this line of work has proposed solutions to some challenges met at PwC in financial NLP. This line of work establishes the ground for continued research and development in multilingual financial NLP and has provided the industrial partner with some contributions tailored to their needs for data resources, model development and evaluation.

3.4 CONFLICT OF INTEREST

The PhD project is built on a collaboration agreement that defines the relationship between the university, company, and PhD student. The PhD student's education and attainment of the PhD degree are given priority over other considerations, and no parties have a direct financial interest in results produced during the PhD. The parties have an interest in expanding their knowledge in machine learning through research, and the company wishes to utilize the knowledge resulting from the research. All four papers of the thesis fulfill the scientific and ethical requirements of the publication venues. The PhD project complies with PwC's requirements for review and permission, where all internal processes are carefully followed. The research work and knowledge are presented to the community in the form of peer-reviewed papers [99–102]. The produced datasets in chapters 5, 6 and 7 are publicly available. The MDAPT models are also publicly available.

Part II

MACHINE LEARNING FOR FINANCIAL TRANSACTION CLASSIFICATION

4 MACHINE LEARNING FOR FINANCIAL TRANSACTION CLASSIFICATION ACROSS COMPANIES USING CHARACTER-LEVEL WORD EMBEDDINGS OF TEXT FIELDS

The following chapter is based on the article "Machine Learning for Financial Transaction Classification across Companies using Character-Level Word Embeddings of Text Fields." by Rasmus Kær Jørgensen and Christian Igel, published in *Intelligent Systems in Accounting, Finance and Management* 28.3 (2021), pp. 159–172 [102]. Appendix A.1 contains supplementary information for this chapter.

ABSTRACT

An important initial step in accounting is mapping financial transfers to the corresponding accounts. We devised machine learning based systems that automate this process. They use word embeddings with character-level features to process transaction texts. When considering 473 companies independently, our approach achieved an average top-1 accuracy of 80.50%, outperforming baselines that exclude the transaction texts or rely on a lexical bag-of-words text representation. We extended the approach to generalizes across companies and even across different corporate sectors. After standardization of the account structures and careful feature engineering, a single classifier trained on 44 companies from 28 sectors achieved a test accuracy of more than 80%. When trained on 43 companies and tested on the remaining one, the system achieved an average performance of 64.62%. This rate increased to nearly 70% when considering only the largest sector.

KEYWORDS: ACCOUNTING; FINANCE; FINANCIAL TRANSACTIONS; RANDOM FOREST; WORD EMBEDDING; MULTICLASS CLASSIFICATION

4.1 INTRODUCTION

Accounting and bookkeeping are essential for every company. They are required by international corporate law, and the process is characterized by repetitive and standardized tasks. Accordingly, there has been a long-standing interest in automating accounting tasks. Mui and McCarthy [148] have already described the use of artificial intelligence (AI) techniques in financial decision-making and concluded that AI methods are promising in this domain. However, despite the digitization of companies, accounting and bookkeeping are mainly carried out by humans, not machines [19].

Date	Transaction Text	Amount	Account Code
17/01/19	DK0477 item xzy 777	799	100 (Sales)
24/01/19	XYZ.COM*StoreName 9376	-1048	230 (Supplies)
08/02/19	BS-123 Housing Company	-9943	210 (Rent)
14/01/19	budgettrans.-10476	385	270 (Utilities)
19/02/19	DK2548 fabric store A/S	-8586	230 (Supplies)

Table 4: Simulated examples of financial transactions and their mapping to account codes. These transactions are not from an existing company, but are representative for the real-world data from Danish small-to-medium sized companies considered in this study.

Many companies commission accounting firms to handle their accounting and bookkeeping for them, simply providing access to transaction data, documentation, and other relevant information. One of the most frequent accounting tasks is the mapping of the daily financial transactions to accounts. In this study, we consider machine learning systems for supporting accounting firms in doing this mapping.

Accounts are units that record and summarize a company's transactions. Companies have several accounts that together compose a hierarchy described by the *chart of accounts*. Table 4 shows examples of transactions and corresponding accounts (codes), and Table 5 exemplifies a chart. The examples in the tables comprise simulated data, in that they do not correspond to an existing company, but they are representative for the real-world data from Danish small-to-medium-sized companies considered in this study. The chart of accounts segregates expenditures, revenues, equity, assets, and liabilities into categories providing a structured overview of the company's finances, which is reported in financial statements of the company. In general, the chart of accounts is not standardized, and a company is free to design a list of accounts as long as the financial reporting follows the regulations and laws. This makes automatic processing difficult. There are both national and international initiatives for standardizing charts of accounts; see Jorge et al. [98] and EUROSTAT [65] for overviews. In 2008, the European commission provided a report with recommendations and good practices on accounting systems for

Account Code	Account Description
Revenues (100-199)	
100	Sales
...	...
Expenses (200-299)	
210	Rent
230	Supplies
270	Utilities
...	...
Assets (300-399)	
330	Inventory
360	Equipment
...	...
Liabilities (400-499)	
400	Tax
440	Interest
...	...
Equity (500-599)	
500	Capital
...	...

Table 5: Simulated simple chart of accounts.

small enterprises, where using a standard chart of accounts was recommended [68]. There are already countries using unified charts [65, 98]; for example, Sweden's BAS chart of accounts [156]. However, the data underlying this study stems from different companies with different charts of accounts.

The task of mapping transactions to accounts is often partly automated by simple rule-based systems, which are specific for each company, often lack accuracy, and require maintenance by experts. What accounting firms need is a system that autonomously learns (from limited training data) to accurately map transactions. Ideally, there would be a system that learns to classify transactions across companies and corporate sectors, even generalizing to new companies for which little or even no historical data exist. In the following, we present such systems. First, we consider a base system that solves the accounting task for each company individually. Second, we generalize this system to handle transactions across different companies. The development of the systems was driven by the following hypotheses:

- Random forests [26] are well suited for learning the mapping from transactions to accounts and should be preferred over simple linear classifiers and nearest neighbor approaches.

- The accuracy of the machine learning approaches increases if text fields in the transactions are included in the analysis using recent natural language processing (NLP) methods [22, 81, 104].

- By using a unified chart of accounts and *transfer learning*, a system can be trained that generalizes over different companies, that is, can be applied to new companies for which only few examples to learn from are available.

To evaluate our systems, we considered real-world transactions from Danish small-to-medium sized companies. We assume that the systems will not operate fully automatically. They will provide suggestions which are then approved by accountants. We considered two measures in our evaluation. First, we evaluated the accuracy of the systems if they would operate fully autonomously and measured how often they predict the correct account (top-1 accuracy). In addition, we considered the case where the systems suggest five accounts to the human expert, from which the expert can efficiently choose. Here we measured how often the right account was among the suggested five (top-5 accuracy).

Our base system, also referred to as **Scenario I** considers accounting financial transactions for a single company. Here it is assumed that only data from that company is available for designing the system. In the development of the base system, we put a focus on character-level embeddings of transaction free-form text fields. In this scenario, the output space (i.e., the set of account codes) is company-specific. The input features are restricted to basic bank transactions containing date, amount, and transaction text. This makes the information in the text field particularly important for accurate classification.

Based on the results from the previous setting, we develop a system for accounting financial transactions across companies and corporate sectors, a setting referred to as **Scenario II**. This allows the system to solve the ”cold-start problem”; that is, to be applicable to companies for which no historical data are available. The first challenge to be met when designing a classifier that can be applied to transactions from several companies is to deal with different charts of accounts. Different charts lead to different output spaces, which complicates the use of machine learning [20]. We address this problem by mapping the individual charts of accounts to a unified chart. Next, one needs to find a representation of the input data that allows for highly accurate classification and generalization across companies. We use our insights from the first scenario and additionally show how to engineer suitable features by combining the transactions with additional data about the involved companies, mimicking the use of background knowledge by accounting and bookkeeping clerks.

The machine learning systems evaluated in the two scenarios rely on the same basic technologies. They demonstrate the feasibility of automating lower levels of accounting processes using machine learning and can be regarded as a significant step towards reaching this goal.

4.2 RELATED WORK

Reviewing the industrial stands and research on AI for processing transaction level operations reveals a consensus that machine learning will be the successor of the common hand-crafted rule-based methods [134, 148, 149, 234]. Still, none of the dominant vendors in this area offers accounting solutions to automate the process entirely. Being a commercial application in a competitive market makes it difficult to identify the classification methods in use and their performance. Bergdorf [20] pointed out that none of today's accounting systems are fully automated and they can only be considered partial solutions at best, since they merely present an improvement to systems using a rule-based approach. They fail to address the problem of standardization that currently hinders the next stage of automation [20].

Most of today's decision support systems for classifying financial transactions are rule-based systems, where the rules are defined by human experts taking into account historical data. When a new transaction matches a rule, it is either assigned to the corresponding account or suggested to the clerks. Transactions not matching any rule are processed manually. The rule-bases require substantial maintenance, because they often require adaptation to a changing environment. Machine learning based approaches allow for autonomous adaptation and promise higher classification accuracies.

Bengtsson and Jansson [19] studied machine learning algorithms for a Swedish accounting system. They evaluated the performance of support vector machines and a feed-forward neural network against the SpeedLedgers[1] implementation of a simple deterministic classifier. Although the results were promising, they did not outperform the existing system. Bergdorf [20] assessed machine learning methods for assigning account codes to invoices and highlights that a problem in automating the process is the need for a more standardized framework, since bookkeeping can be subjective and companies perform it differently. He suggests building a classifier for each known company and to apply a general set of rules based on standard cases for new companies. Bergdorf [20] also reflected on the idea that companies can be widely different, although classifiers for similar companies can benefit from each other.

It has been explored how additional information can be linked to transactions to support classification [20]. Folkestad and Vollset [71] carried out a study on automatic classification of bank transactions in collaboration with the Norwegian SpareBank1, which have used a manual filter to classify transactions to budget categories. That study used external company data to improve the classification. The classification system was built on a bag-of-words (BoW) approach and logistic regression. They found that enriching bank transactions with external company information improved the classification system. In another study, Folkestad et al. [70] investigated the effect of linked

[1] https://www.speedledger.se

open data, using Wikidata and DBpedia, to aid in the classification of bank transactions. They observed that, usually, a company name is present in the transaction text, which suggests finding the company in Wikidata and DBpedia and to retrieve information about what industry the company operates in. However, an adverse effect was found when using the extracted data. Skeppe [206] conducted a study on the classification of Swedish bank transactions with early and late fusion techniques with the goal to improve the classification of bank transactions. The study did not show significant improvements compared to the bank's rule-based system, but still concluded the classification of transactions is well suited for machine learning [206].

When considering natural language information linked to transactions, the question of how to best exploit the highly unstructured information in the text fields is crucial for achieving a performant system. O'Leary and Kandelin [157] stress the importance of NLP for linking transactions to accounting activity. Typically, simple pattern matching or BoW approaches have been applied to vectorize the free-form text fields [19, 70, 71, 225]. Ideally, one would like to use domain specific systems; that is, NLP systems adapted tailored towards accounting language [157]. However, in particular for low resource languages, not enough domain specific training data may be available for building a machine learning based NLP system without using data from other (general) domains and perhaps even other languages.

None of the aforementioned studies explicitly targets the *multiple charts of accounts problem* of companies having different account categories. They have either considered a separate classifier per organization, which means classifiers for different companies cannot share information during training and a new classifier must be built for every new company, or presumed a single predefined chart [19, 20]. Generalization across companies and corporate sectors can be viewed as a domain adaptation problem. Machine learning algorithms typically rely on the assumption that the training and test data are drawn from the same underlying distribution. However, this assumption does not hold for many real-world applications. In practice, there are many cases where the training sample and test sample are from different distributions; for example, when classifying new company transactions using historical data from different companies. Thus, the problem of building a classifier for financial transactions evaluated on unseen companies, and perhaps even on a new corporate sector, can be considered as a *transfer learning* or more precisely a *domain adaptation* task [33, 111, 159]. When testing on new companies, the distribution of the financial transactions changes between training and testing. As we have potentially many companies to learn from, we are dealing with a multi-source domain adaptation task [159, 216].

4.3 REPRESENTING FINANCIAL TRANSACTIONS

This section introduces the methods used for transaction representation and the unification of the output space (i.e., the charts of accounts). Then, the prediction models considered are briefly presented.

4.3.1 *Transaction Text Embeddings*

Current accounting systems rely on counting-based, lexical representations or simple pattern matching to retrieve information from the text fields. However, the unstructured free-form text limits the performance of these approaches. Thus, when classifying transactions, the key question is how to exploit the natural language text. Natural language processing was already used in Prolog based expert systems for automating accounting tasks [158]. Most recent studies apply simple BoW approaches to vectorize the free-form text fields [19, 70, 71, 225] or do not describe the processing of the text features in detail [20, 206]. We address these challenges by using a character-level word representations that exploits sub-word information to represent unseen words as the sum of their character n-grams.

Character-level word embeddings create a low dimensional representation of sequences of words. Compared to simple pattern matching, this produces a more refined distance measure between the sequences of words that fosters generalization. This distance is used to identify how similar transactions are based on the features extracted from their textual information. Several software solutions are available to compute word embeddings, such as Doc2Vec and Word2Vec [143, 144] and GloVe [162]. Because of the limited training data we need to rely on out-of-domain text sources. In addition, our data requires support of the Danish language, which is used in the transaction text. Therefore, we decided to use fastText, which offers pretrained models for 157 languages including Danish [22, 81, 104]. These models were trained on data from the Common Crawl Project and the free online encyclopedia Wikipedia [81] using CBOW [143] (with position-weights, output dimension 300, character n-grams of length 5, a window of size 5, and 10 negatives). The main difference between fastText and both Word2Vec and Glove is the use of the smallest n-gram unit. Both GloVe and Word2Vec treat each word in the corpus as the smallest unit. fastText recognize each word as composed of character n-grams and treats each character n-gram as the smallest unit. Therefore, GloVe and Word2Vec only learn vectors for the words contained in the training corpus and cannot construct a meaningful vector for out-of-vocabulary words. fastText characterizes words as the sum of their character n-grams and the word itself, if the word is in the vocabulary. The option of computing a representation through summarizing character n-grams makes it possible to construct representations of out-of-vocabulary words (i.e., improvised words, abbreviated words or misspelled words) and to identify their closest neighbors among the words in the vocabulary. This is important when dealing with free-form text fields in transactions. fastText constructs a word representation, and we use the straight-forward method of creating a sentence representation for a text field by averaging the word embeddings [105].

PRE-PROCESSING We apply standard pre-processing steps such as tokenization and lowercasing the corpus. Other common operations such as spelling corrections, stemming and lemmatization are omitted. Transforming a word into its word stem is not helpful in our case. Given the nature of transactions, knowing whether the transaction represents commerce of one item (singular) or several items (plural) can be informative. We replace common abbreviations in accounting by the corresponding natural language expressions. We also remove stop-words, punctuation, digits and non-alphanumeric characters.

LOW DIMENSIONAL TRANSACTION TEXT EMBEDDING We perform a principal component analysis (PCA) to reduce the feature space of the averaged word embeddings from the standard output dimensionality of fastText $\mathbb{R}^{300}$ to $\mathbb{R}^{d_{PC}}$ for a small number of components d_{PC}. An example of the text feature generation is given in Table 6.

(1) Raw transaction text	["PS-DK9988-776655 internet"]
(2) Pre-processed text	["payment service", "internet"]
(3) Initial vector representation	$[0.8, 0.5, \dots, 0.2, 0.4, 0.7] \in \mathbb{R}^{300}$
(4) Final representation after PCA	$[0.1, \dots, -0.2, 1.7] \in \mathbb{R}^{d_{PC}}$

Table 6: Example of the representation of text features.

4.3.2 *Unifying Chart of Accounts*

In this work, we consider two scenarios. Section 4.4 considers companies individually; that is, each company represents a separate data set with its own specific label space. However, in Section 4.5, we study companies collectively; that is, each company represents a different, but related data set with shared label space.

In the second scenario, we need to address the *multiple chart of accounts problem* and define the shared label space for classifying the transactions across entities. Different companies have *Individual Charts of Accounts*. That is, corresponding accounts of two companies may have different account codes. To train a classifier that generalizes across companies, we manually map the individual charts of accounts to a single *Unified Chart of Accounts*. We examined several existing templates, none of which appeared to perfectly suit for our task that considers various corporate sectors. The template that best fit the companies is a standard chart designed for so-called *Danish Class A* companies. The template is for smaller companies, which fits most companies in this study. It consists of 194 accounts, where 80 are revenue and expenditure accounts. Not all of them are used by the companies in our study.

4.3.3 *Classifiers*

We use a standard *random forest* for classification [26]. In general, random forests give good results in practice [69], the possibility to compute the *Out-of-Bag* (OOB) error, and the robustness with respect to hyperparameters. This make model selection comparatively easy, and feature selection is handled by design.

To evaluate the performance of the random forest on the classification problem, we consider three baseline models: *logistic regression*, k-*nearest neighbor*, as well as a *majority class classifier,* predicting simply the most frequent class in the training sample. Thus, we have baselines from a linear parametric model, a non-linear non-parametric model, and a majority class classifier that can be regarded as the trivial baseline.

4.4 SCENARIO I: INDIVIDUAL ACCOUNT CHARTS

We started by looking at the scenario predominantly considered in the literature, in which a single company is considered and it is assumed that only historic data from this company is available. The main goal of this part of the study was to show the advantages of the proposed way of representing the textual transaction information.

4.4.1 *Data*

The data for this study was collected from 473 Danish small to medium-sized companies, yielding a total of 313,878 financial transactions. We transformed the temporal information, and describe each transaction by the following $d_{PC} + 5$ features: Amount, Week, Month, Quarter, Year, and the d_{PC} text features. The label space is defined by the entire chart of accounts, as exemplified in Table 5. The label space of the individual companies ranged from a minimum of two classes (meaning, the company only used two accounts) to a maximum of 513 classes with 52 classes on average.

4.4.2 *Experimental Setup*

The task was to predict the accounts of future transactions given labeled data from previous transactions. For each company, we ordered the financial transactions by time. The data was then partitioned using the first 70% of the company's financial transactions for training, and the subsequent 30% (89,130 transactions) for testing (classes in the test set not included in the training set were removed). The setting is depicted in Figure 1. We chose this setting because it is closest to reality: The system is built based on data from the past and applied to future data. However, the transaction generating process cannot be assumed to be stationary, which poses a challenge that all algorithms considered have to cope with. This should be taken into account when evaluating the classification accuracies on the test data. Randomizing the data independent of time would lead to overoptimistic performance estimates.

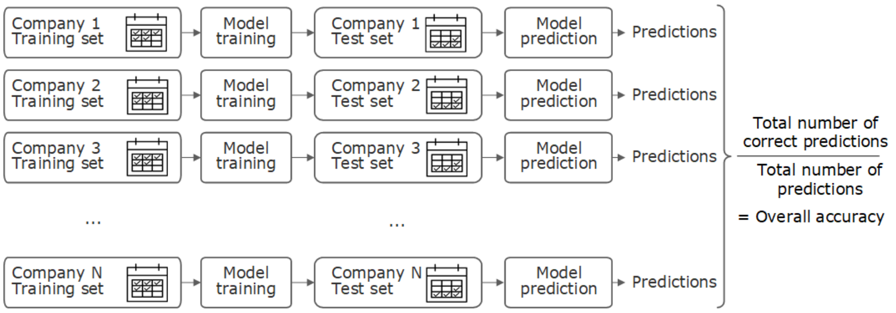


Figure 1: Scenario I: *Transaction Time Series setting* for individual companies.

For each experimental setup, we built an independent classifier per company. During testing, each of these 473 classifiers was applied to the test transactions from the corresponding company. The average top-1 classification accuracy was reported over the resulting 89,130 predictions. Each experiment was conducted twice, using either the fastText or the lexical BoW embedding. Furthermore, an ablation study was performed to investigate the importance of the features extracted from the text fields.

The performance of the random forest classifier was studied for all companies in three settings:

A using all features.

B only the transaction text features.

C all features except the transaction text features.

In addition, we computed the top-5 accuracy using the probabilistic output of the random forest in the best settings. For the top-5 accuracy, we only considered companies with more than 100 classes. This subset consists of 58 companies with 160 classes on average and 87,205 transactions.

The number of trees in the random forest was set to 300. For experiments with less than 300 training data points, the number of trees was changed to be two-thirds of the training set size. The tree depth was not restricted. Growing trees to their full extension allows us to capture the rare classes [76]. The number of candidate variables for splitting was set to the squareroot of the input dimensionality $\sqrt{d_{PC}+5}$. When not using a fixed value of 10 principal components, which was chosen after inspecting the eigenspectra of the training data [72], we used the OOB training error also to adjust the number of principal components for the word embedding by choosing $d_{PC} \in \{2^i \,|\, i = 0, \dots, 7\}$ for each company individually.

The character-level text embedding was compared to a lexcical BoW approach (using term frequency–inverse document frequency, TF-IDF) for each company. For the BoW, the same preprocessing was applied, except that stemming was added. Words that occurred less than three times were removed (except for two companies with little training data). Similar dimensionality reduction operations as used for the character-level word embeddings were performed on the resulting word-count vectors in the training set to reduce the feature dimensionality.

4.4.3 *Results*

The results for predicting transactions from the 473 companies are summarized in Table 7. In all experiments, the random forest classifiers outperformed the baseline methods.[2] In the random forest experiments using text features (A and B), the character-level text embedding gave significantly better results than the lexical BoW approach did (Pearson's chi-squared test, $p < 0.001$); therefore, we restrict our discussion to the character-level text embedding in the following. Choosing the number of PCA components for the random forest using the OOB error gave slightly better results ($p < 0.001$).

Methods	**Overall**	**A**	**B**	**C**
Majority Class	$\text{Accuracy}^{10}_{\text{BoW}}$	32.65%	32.65%	32.65%
	$\text{Accuracy}^{10}_{\text{TTE}}$	32.65%	32.65%	
Logistic Regression	$\text{Accuracy}^{10}_{\text{BoW}}$	20.51%	60.03%	20.11%
	$\text{Accuracy}^{10}_{\text{TTE}}$	20.31%	68.11%	
k-Nearest Neighbor	$\text{Accuracy}^{10}_{\text{BoW}}$	49.34%	68.75%	40.04%
	$\text{Accuracy}^{10}_{\text{TTE}}$	47.84%	71.34%	
Random Forest	$\text{Accuracy}^{10}_{\text{BoW}}$	76.87%	71.88%	**45.75%**
	$\text{Accuracy}^{10}_{\text{TTE}}$	*77.33%*	*76.32%*	
	$\text{Accuracy}_{\text{BoW}}$	77.73%	73.56%	
	$\text{Accuracy}_{\text{TTE}}$	**80.50%**	**77.29%**	

Table 7: Accuracies for complete feature set (A), only Transaction Text (B), and without Transaction Text (C). Transaction Text Embedding and the BoW approach are denote by $\text{Accuracy}_{\text{TTE}}$ and $\text{Accuracy}_{\text{BoW}}$, respectively. The superscipt 10 refers to using a fixed number of 10 components for the vector representation; no superscript indicates that the number of components was selected using the OOB training error. Results are averaged over all considered companies. Best results are marked with bold.

The results of experiment A using all features and individual adjustment of the number of principal components show that 80.50% of the test transactions could be classified correctly by the random for-

[2] For k-nearest neighbor we present the result for the best k selected on a coarse grid.

est. In the top-5 setting, 86.57% of 23,744 test transactions are within the recommendations.

An ablation study was conducted to evaluate the importance of the input features. In setting B, only the transaction text features were provided as inputs, and the random forest could still obtain a test accuracy of 77.29%, only 3.21 percentage points (p.p.) less in comparison to A using the complete feature set. In experiment C, all features except the transaction text features were provided, leading to a drop in accuracy to 45.75%, a decrease of 34.75 p.p. compared with A.

4.5 SCENARIO II: UNIFIED CHART OF ACCOUNTS

Building on our results from the first scenario, we devised a system for classifying transactions that works across companies and even across different corporate sectors. In the following, we first describe the preprocessing of the data. This includes combining information from different sources as well as nonstandard data preprocessing, resulting in 28 features (see Table 10) per financial transaction. Then, we discuss the application of the unification process introduced in Section 4.3.2. Finally, we present the results from our evaluation of the resulting system in various settings.

4.5.1 *Data Collection and Preprocessing*

The data was collected from 44 Danish small to medium-sized companies, referred to as the *subject companies*. They transact with a vast number of trade partners, referred to as the *external companies*. The analysis was restricted to domestic operating activities related to revenue and expenditure accounts, documented by invoices. The 44 subject companies represent 28 sectors. A total of 10,354 transactions was considered.

4.5.1.1 *Data Fusion and Preprocessing*

Accounting and bookkeeping clerks do not solely rely on the information in the journal records containing the transactions, but often base their decisions on experience and additional domain-knowledge. We studied their decision making process and made the following observations. First, when additional support is needed to make a decision, they access the invoice to inspect the financial transaction. Then, they evaluate the companies, the sectors, and the types of service or products they exchange. To match the human performance, we need to access the additional information and provide it as input to the classifier. Therefore, we fused the following data sources:

JOURNAL DATA holds the company's financial transactions in a systematic and chronological order.

INVOICE DATA contains information specifying the exchange of goods or services between the two parties.

COMPANY DATA is retrieved from the Danish state's company register [219].

4.5.1.2 *Basic Feature Engineering*

Approximate string matching was employed to derive a discrete *Payment Type* feature from the transaction text. We created a list of common transaction keywords ("credit card", "bank transfer", "automatic payment service", "fees" or "unknown type") and associated them with payment type categories. Subsequently, we used the Levenshtein distance to measure the similarity between the transaction text and the keywords. The payment type of the closest keyword determined the Payment Type feature.

An often informative feature is the time interval between the date of issue and payment. The difference between the date of issue and the date of payment may carry information about the complexity and arrangement of the transaction. Therefore, we computed a *Difference in Days* feature measuring this time interval.

Another descriptive feature is the geolocational information. For deriving a feature for the *Distance* between the two entities, we examined the distinctive locations of the transacting companies. The structure of companies is often divided into several subsidiary companies, which implies different addresses. By calculating the distance between the subject company and the external company, the different subsidiaries result in distinct distances. This feature may capture characteristics that cannot be inferred from the high-level names of the companies. For distance computation, latitude and longitude coordinates of external companies and subject companies were extracted using a geocoding web service. The haversine formula was used to compute the distance between the two locations. It calculates the *great-circle* distance over the earth's surface and implements a stable distance measure [94]. Let (φ_1, λ_1) and (φ_2, λ_2) denote the latitude and longitude of two transacting entities; then, with $\nabla\varphi = \varphi_2 - \varphi_1$ and $\nabla\lambda = \lambda_2 - \lambda_1$, we compute

$$a = \sin^2\left(\frac{\nabla\varphi}{2}\right) + \cos(\varphi_1)\cdot\cos(\varphi_2)\cdot\sin^2\left(\frac{\nabla\lambda}{2}\right). \tag{1}$$

The great circle distance is described by $c = 2\,\mathrm{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$, which corresponds to a distance $d = R \cdot c$ between the two locations, where R is the earth's mean radius (≈ 6371 km).

4.5.1.3 *Transaction Text and Sector Text Features*

We derived three text representations from each transaction: *Transaction Text* as well as *Subject Sector Text* and *External Sector Text* with the two latter describing the sectors of the subject company and the external company, respectively. Transaction Text and the two sector texts contain valuable information for discerning transactions. To derive features that support generalization across companies, we used the character-level word embeddings introduced in Section 4.3.1 and giving good performance as shown in Section 4.4 to create a low dimensional representation of the sequences of words. Training on the available accounting data could improve the word embedding, but bears some risk of overfitting. Thus, we used a generic word embedding for the main language of the country in which the transactions were obtained. Studying domains specific embeddings and multi-lingual embeddings is left to future research.

Following the creation of the averaged word embeddings for the three text fields, we obtained three 300-dimensional feature vectors. For each of the three text features, we perform a PCA on the training data to reduce the feature space of the averaged word embeddings from $\mathbb{R}^{300}$ to $\mathbb{R}^{5}$. That is, we obtain five-dimensional feature vectors for the Transaction Text, Subject Sector Text and External Sector Text, leading to 15 additional features per transaction in total.

4.5.1.4 *Unifying Charts of Accounts*

We manually converted the individual charts to the Unified Chart of Accounts, which then defines the class labels. Experienced accounting and bookkeeping clerks assisted in this process. Most of the conversions were carried out by the same clerks who are responsible for the accounting and bookkeeping of these companies. Comprehensive instructions were communicated to the clerks before the conversion process. Consensus among the converters was required. If consensus could not be reached, the account was not converted. The strict procedure minimizes errors and the risk of introducing biases in the ground truth data. For the companies we considered, 69.67% (3,549) of the accounts were mapped to an account in the Unified Chart of Accounts, leaving 1,545 accounts unmapped. The process reduced the 3,549 different accounts to 54 out of the 194 accounts in the Unified Chart of Accounts, resulting in a 54-class classification task. Not all companies used all 54 classes. Figure 2 shows the number of account codes used by the different companies.

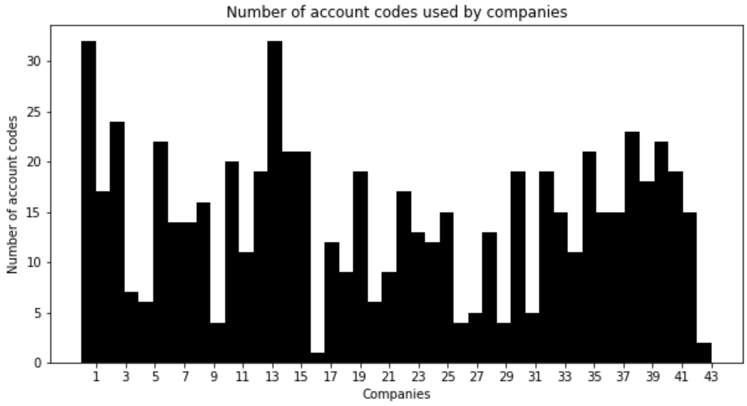


Figure 2: Number of account codes used by companies.

That not all accounts could be converted was expected and is less of a problem than it may appear. First, many of the unmapped accounts were inactive accounts; and if they were not inactive, they were rarely used. Second, some accounts contained non-specific financial transactions that suffered from too much customization and needed correction. Third, some accounts were specialized types of accounts with company-specific financial transactions. In this study, we left these accounts unmapped. In practice, one can optimize the companies' structure of accounts to deal with the first two cases and extend the Unified Chart of Accounts to deal with the latter case.

4.5.2 *Experimental Setup*

We empirically evaluated our approach in two different general settings. Setting A is referred to as the *Transaction Time Series* setting, where we trained on older data of all companies and evaluated the system by predicting more recent data of these companies. Setting B, called *Classification of New Company Transactions*, used all data from a subset of the companies to classify the transactions from companies not in the training set. We build on the results from our previous study and used the same hyperparameter settings as in Section 4.4. We did not vary the number d_{PC} of components and set the number of candidate variables for splitting to $\sqrt{d}$.

TRANSACTION TIME SERIES (SETTING A) The first general setting considered the scenario in which we have historical labelled data and want to predict the accounts of future transactions. For each company, we ordered the financial transactions by time. The data were partitioned similar to *Individual Account Charts* experiments by using the first 70% of a company's financial transactions for training and the subsequent 30% for testing. As we want to build a single classifier for all companies, we combined the training and test data of all 44 companies to a joint training and a single test set. Figure 3 illustrates this setting. In this case, the classifier has to generalize across 28 different corporate sectors. Furthermore, we considered the simpler scenario in which the classifier has to generalize only across companies of a single corporate sector. In this, we considered only the 6 companies of the largest sector in our data set: pharmacies. The pharmacy sector contains 3,262 training examples, 1,394 test examples and 25 classes. In extension, an ablation study was performed to investigate the importance of the features extracted from the text fields.

In summary, we considered the following five scenarios in the general setting A:

A.I All features and all companies were included.

A.II All features were provided, but only pharmacies were considered.

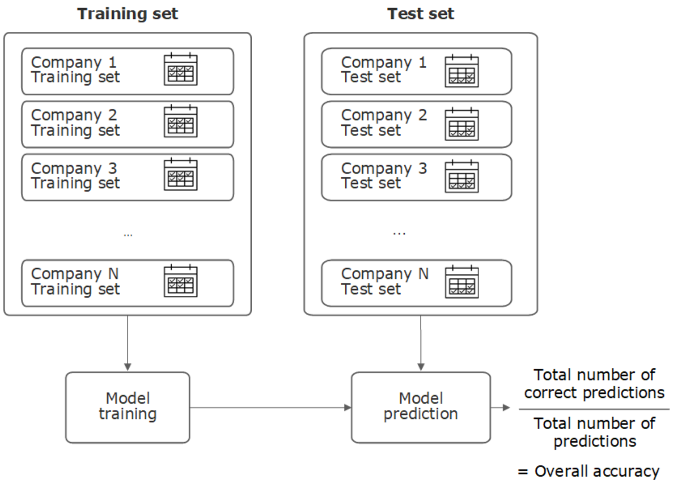


Figure 3: Scenario II: Transaction Time Series (setting A) for several companies.

A.III Only the Transaction Text features were provided and all companies were included.

A.IV All features except the Transaction Text features were considered and all companies were included.

A.V All text field features were excluded and all companies were considered.

CLASSIFICATION OF NEW COMPANY TRANSACTIONS (SETTING B) The second general experimental setting models the real-world scenario of classifying financial transactions of new, previously unseen companies when no historical labeled transactions are available. The main questions we wanted to answer were: Can a classifier built using our approach generalize to a new company? Can it generalize to a new corporate sector? Does the task become easier if we restrict training and testing to companies from a single sector, that is, consider less but more homogeneous data?

We considered two experimental settings to answer these questions. We trained one classifier on all available companies across all corporate sectors. Next, we trained a classifier only on the training data from the same sector as the new company. The first approach has the advantage that the classifier uses more training data. The second approach uses training data that is more similar to the data at test time, which may prevent wrong generalization. Furthermore, we considered an even more difficult generalization task, namely to generalize to several companies from a completely new corporate sector. All experiments used the full feature set:

B.I *Leave-One-Company-Out:* All subject companies except one were used for training the classifier, which was tested on the left out company. The results were averaged over the 44 possible splits into training and test companies.

b.ii *Leave-One-Sector-Out:* Companies were grouped according to their respective corporate sectors. All subject companies from all sectors except one were used for training the classifier, which was then tested on the companies from the hold out sector. The results were averaged over the 28 possible splits into training and test companies.

b.iii *Leave-One-Pharmacy-Out:* We considered only the corporate sector for which we had most transactions, the pharmacy sector. The data from all pharmacies except one are used for training the classifier, which was tested on the transactions from the hold out pharmacy. The results were averaged over the 6 possible splits into training and test companies.

The data handling in the leave-one-out settings is illustrated in Figure 4.

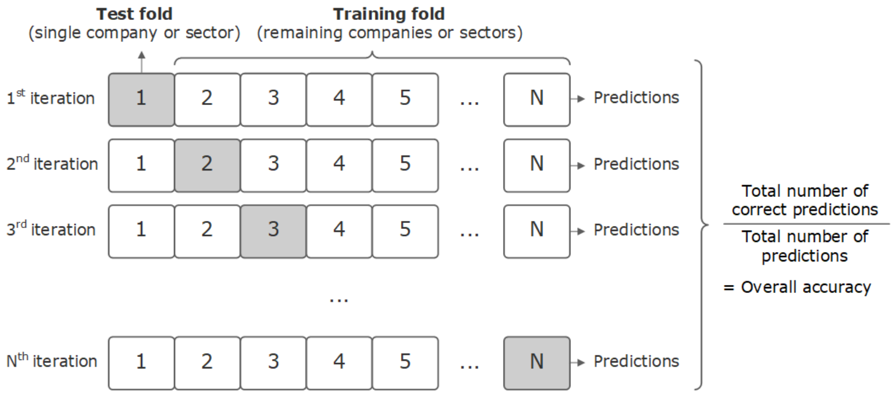

Test fold
(single company or sector)
Training fold
(remaining companies or sectors)
1st iteration
2nd iteration
3rd iteration
Nth iteration
1
2
3
4
5
...
N
Predictions
Total number of correct predictions
Total number of predictions
= Overall accuracy

Figure 4: Scenario II: Hold-One-Out setting for Classification of New Company Transactions (setting B).

As discussed in Section 4.3, training on some companies and generalizing to other companies can be viewed as a multi-source domain adaptation problem [159, 216]. Settings B.I and B.II addresses the domain adaptation task by an *aggressive approach* [21], which combines all available data (all companies) into one source domain. This approach can profit from a large training data set. In contrast, setting B.III follows a *selective approach* that only combines selected sources, the within-sector companies, into a source domain. In this case, we expect source and target distribution to be more similar than in aggressive approaches [21]. This may reduce potential negative transfer that can occur when data from the source distribution has a negative impact on the performance on the target distribution; however, it requires enough training data for the individual classifiers [21, 216]. Because of the latter, experiment B.III considered only the pharmacy sector for which we had most training data.

Data Setting	All Companies	Subset of data: Pharmacy Sector
Feature Setting	Complete Feature Set	Complete Feature Set
Experiment	A.I	A.II
Majority Class	23.82%	18.51%
k-Nearest Neighbor	37.74%	47.35%
Logistic Regression	44.99%	26.83%
Random Forest	**80.76%**	**81.50%**

Table 8: Results of Transaction Time Series experiments using complete feature set. All results refer to (average) test errors. The k-nearest neighbor results always refer to the best test accuracy for $k \in \{1, \ldots, 10\}$. Best results are marked with bold.

Data Setting	All Companies	All Companies	All Companies
Feature Setting	Only Transaction Text	Without Transaction Text	Without Text features
Experiment	A.III	A.IV	A.V
Majority Class	23.82%	23.82%	23.82%
k-Nearest Neighbor	67.16%	37.70%	37.48%
Logistic Regression	36.02%	47.53%	29.27%
Random Forest	**69.63%**	**78.09%**	**56.72%**

Table 9: Results of the ablation study for the Transaction Time Series experiments. All results refer to (average) test errors. The k-nearest neighbor results always refer to the best test accuracy for $k \in \{1, \ldots, 10\}$. Best results are marked with bold.

4.5.3 *Results*

4.5.3.1 *Transaction Time Series (Setting A)*

CLASSIFICATION PERFORMANCE USING ALL FEATURES In all experiments, the random forest classifiers clearly outperformed the baseline methods; see Table 8 and 9. The results of the basic experiment A.I in Table 8 show that 80.76% of the test transactions could be classified correctly by the random forest. In experiment A.II, we only considered the companies from the largest sector. The random forest achieved a test accuracy of 81.50%. This is only slightly better than the average in A.I. Still, it indicates the system can profit from building individual classifiers for specific corporate sectors if enough data from the sector are available.

FEATURES IMPORTANCE To evaluate the general importance of the input features, we analyzed the random forest model found in experiment A.I. We determined the relative importance of the input variables by the mean decrease impurity metric [27]. Table 10 shows the feature importance from A.I in decreasing order. The relative feature importances show that the decisions by the random forest strongly rely on the text features. Also the derived features Distance and Difference in Days are ranked higher than most basic features. The ab-

Importance	Feature
0.084126	External Sector Text 1st Comp.
0.076574	Transaction Text 1st Comp.
0.071125	External Sector Text 2nd Comp.
0.065961	External Sector Text 4th Comp.
0.063809	External Sector Text 3rd Comp.
0.062806	External Sector Text 5th Comp.
0.056514	Transaction Text 2nd Comp.
0.053486	Transaction Text 4th Comp.
0.052116	Transaction Text 3rd Comp.
0.047529	Amount
0.045169	Transaction Text 5th Comp.
0.038887	Distance
0.036161	VAT Amount
0.027632	Subject Sector Text 1st Comp.
0.027476	Difference in Days
0.024845	External Business Entity
0.021755	Subject Sector Text 4th Comp.
0.021032	Payment Type
0.019611	Subject Sector Text 5th Comp.
0.018122	Subject Sector Text 2nd Comp.
0.018101	Subject Sector Text 3rd Comp.
0.017512	Issued Week
0.016711	Paid Week
0.011049	Subject Business Entity
0.006678	Paid Quarter
0.006569	Issued Quarter
0.005313	Paid Year
0.003329	Issued Year

Table 10: Feature Importance (A.I).

lation studies provided further insights. Using only the Transaction Text features (A.III) to classify financial transactions for the 44 companies gave a test accuracy of 69.63%. This shows the random forest can retrieve sufficient information just from the Transaction Text to map financial transactions to account codes with a high accuracy. In comparison to the setting with the complete feature set, we observed a decrease of 11.13 p.p..

In experiment A.IV, all features except the Transaction Text features were provided. The random forest obtained a test accuracy of 78.09%, a decline of 2.67 p.p. in comparison to using the complete feature set. When we excluded all text features in experiment A.V, the test accuracy dropped to 56.72%, which is 24.04 p.p. less compared to using the complete feature set. Thus, the Transaction Text and the two sector text features provide information that can be used to classify transactions and that is not contained in the other features. Evidently, it could not be obtained with the original three features available when starting this study (see Table 4).

4.5.3.2 *Classification of New Company Transactions (Setting B)*

Experiments	**Leave-One Company-Out B.I**	**Leave-One Sector-Out B.II**	**Leave-One Pharmacy-Out B.III**
Source Domain	43 companies	27 sectors	5 pharmacies
Target Domain	1 company	1 sector	1 pharmacy
Number of examples	10,354	10,354	4,656
Classification Accuracy	**64.62%**	**51.94%**	**69.95%**

Table 11: Results of Classification of New Company Transaction. All results refer to (average) test errors over the combination of source and target domains. Best results are marked with bold.

The results for predicting transactions from new companies are summarized in Table 11. The *Leave-One-Company-Out* experiment B.I showed an average test performance of 64.62% over the 44 combinations of source- and target domains. Inspecting the results further, we observed good generalization to target domains with one or more within-sector companies in the source domain, but a lower generalization to target domains with no within-sector companies in the source domain. The results indicate the ability to generalize well is linked to the similarities between source and target domain. Comparing Leave-One-Sector-Out with Leave-One-Company-Out in experiment B.II reveals a drop in accuracy to an overall performance of 51.94%. The dissimilarities between the source and target distributions were higher, and the complementary information was lower in comparison to B.I. Furthermore, it turned out to be more difficult to generalize to companies such as pharmacies, manufacturing of jewelry and driving schools compared to cosmetology, grocers and convenience stores. The finding indicates that some companies are more peculiar and niche in their operations than others, and so are the financial transactions from these companies.

The Leave-One-Pharmacy-Out setting (B.III) only considered within-sector companies. Hence, source and target distribution can be expected to be closely related. The experiment gave a test accuracy of 69.95%. Thus, we achieved better generalization by only considering within-sector companies.

4.6 DISCUSSION AND CONCLUSION

FEEDBACK FROM ACCOUNTANTS Albeit the accuracies achieved by the proposed systems in both scenario are high, the question arises how valuable these results are for practitioners when the system is used to assist humans in classifying transactions by making suggestions for assignments. Therefore, we informally presented our study to eight accounting and bookkeeping experts (three senior managers; two senior associates; two directors; one partner) managing accounts from which the data for this study were taken. Their feedback was unanimous and is summarized in the following. Though this summary cannot be regarded as a scientific result, we found it insightful:

- In the evaluation of the experts, the systems presented in experiments scenario I (A) and scenario II (A.I) could approximately save 30-50% time and costs in comparison to their knowledge on rule-based systems and manual approaches.

- The possibility to simply retrain the classifiers to update the system is seen as an advantage. In rule-based systems, outdated mappings can require humans to review the whole rule-base, which is a tedious procedure.

- The solution to the *cold-start problem*, which arises when an accounting firm has to deal with new companies, was very appreciated. Today, a new company means manually mapping all financial transactions to account codes and constructing rules to be implemented in the rule-based system. Experiment B demonstrated the ability to progress from 0% (all manual) to 51.94% – 69.95% (automatically mapped transactions based on data from other companies).

- It was appreciated that the approach can be transferred to other countries. The Unified Chart of Accounts is sufficiently general and there are pretrained word embeddings available in more than 150 languages.

- Because of high accuracies of the systems, they are currently being evaluated by one of the largest international accounting firms.

FUTURE WORK Apart from using more training data, the biggest room for improvement lies in the conversion to the Unified Chart of Accounts. Better templates can be designed and companies could be encouraged to (mainly) use predefined templates. According to the accounting and bookkeeping experts consulted in this study, this would be feasible in the near future.

So far, we used a word embedding trained on generic data from a single language. Studying embeddings trained specifically on accounting language, which results in domain-dependent NLP systems [157], and multi-lingual embeddings is left to future research.

CONCLUSIONS We have presented novel systems for supporting accounting firms in mapping financial transactions to the corresponding accounts. In the first scenario, a highly performant semi-automatic approach to train company-specific classifiers mapping to an individual chart of accounts was developed. It provided top-5 recommendations with an average accuracy of 86.57%. The system devised for in the second scenario can be regarded as the next step towards the development of semi-automatic or even a fully automatic system for processing transactions. The approach allows to generalizes across companies and even to new companies, in contrast to the company-specific classifiers or rules used in industrial systems. Although we had to discard accounts that could not be mapped to the Unified Chart of Accounts defining the label space, we regard test accuracies over 80% when classifying new transactions from known companies and almost 65% for transactions from new companies without historical data as high. As confirmed (in informal, not scientifically controlled interviews) by accounting and bookkeeping clerks who work with our data, these results are superior to the rule-based system they use. Our technical findings are rather general. In particular, the following insights may be valuable for other accounting systems and systems dealing with transaction data in general:

- The study on feature importance, specifically the ablation experiments, showed the discriminative power of the engineered features. The derived features measuring the distance between companies and the time between the date of issue and the date of payment, respectively, were more important for the random forest classifier than most basic features.

- Our main focus was on features generated from free-form text fields. The resulting features Transaction Text and External Sector Text turned out to be highly important for the random forest and the ablation studies. Not providing the text features led to a performance drop of more than 20 p.p., which shows that they provided information that could not be extracted from the other features.

- In contrast to previous studies, we used pretrained, non-task-specific word embeddings with character-level features. This approach provides a better embedding for words not in the vocabulary, which occur frequently in the unstructured and sometimes improvised transaction texts. This is reflected in our experiments, where a lexical BoW approach performed significantly worse. Our text processing does not require training on domain-specific transaction data. This is particularly useful when only little historical data is available.

- We demonstrated that it is possible to classify transactions of a new company without historic data. Across all 28 corporate sectors, we achieved an average accuracy of almost 65% for new companies. When we restricted the study to a single company sector, we reached almost 70%. Thus, the *cold-start problem* can be reduced significantly.

It is widely acknowledged that machine learning based systems for accounting financial transaction will replace software systems relying on handcrafted rule-bases and will increase the degree of automation in this area, and the results of our study support this prediction.

acknowledgments We thank PwC for providing the data and thank the accounting and bookkeeping experts at PwC for their support and valuable feedback.

Part III

MULTILINGUAL NLP

5

MDAPT: MULTILINGUAL DOMAIN ADAPTIVE PRETRAINING IN A SINGLE MODEL

The following chapter is based on the article "mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model." by Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott, published in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3404–3418 [101]. Appendix A.2 contains supplementary information for this chapter.

ABSTRACT

Domain adaptive pretraining, i.e. the continued unsupervised pretraining of a language model on domain-specific text, improves the modelling of text for downstream tasks within the domain. Numerous real-world applications are based on domain-specific text, e.g. working with financial or biomedical documents, and these applications often need to support multiple languages. However, large-scale domain-specific multilingual pretraining data for such scenarios can be difficult to obtain, due to regulations, legislation, or simply a lack of language- and domain-specific text. One solution is to train a single multilingual model, taking advantage of the data available in as many languages as possible. In this work, we explore the benefits of domain adaptive pretraining with a focus on adapting to multiple languages within a specific domain. We propose different techniques to compose pretraining corpora that enable a language model to both become domain-specific and multilingual. Evaluation on nine domain-specific datasets—for biomedical named entity recognition and financial sentence classification—covering seven different languages show that a single multilingual domain-specific model can outperform the general multilingual model, and performs close to its monolingual counterpart. This finding holds across two different pretraining methods, adapter-based pretraining and full model pretraining.

5.1 INTRODUCTION

The unsupervised pretraining of language models on unlabelled text has proven useful to many natural language processing tasks. The success of this approach is a combination of deep neural networks [224], the masked language modeling objective [59], and large-scale corpora [248]. In fact, unlabelled data is so important that better downstream task performance can be realized by pretraining models on more unique tokens, without repeating any examples, instead of iterating over smaller datasets [178]. When it is not possible to find vast amounts of unlabelled text, a better option is to continue pretraining a model on domain-specific unlabelled text [53, 84], referred to as domain adaptive pretraining [83]. This results in a better initialization for consequent fine-tuning for a downstream task in the specific domain, either on target domain data directly [83], or if unavailable on source domain data [84].

The majority of domain-adapted models are trained on English domain-specific text, given the availability of English language data. However, many real-world applications, such as working with financial documents [8], biomedical text [120], and legal opinions and rulings [35], should be expected to work in multiple languages. For such applications, annotated target task datasets might be available, but we lack a good pretrained model that we can fine-tune on these datasets.

In this paper, we propose a method for domain adaptive pretraining of a single domain-specific multilingual language model that can be fine-tuned for tasks within that domain in multiple languages. There are several reasons for wanting to train a single model: (i) Data availability: we cannot always find domain-specific text in multiple languages so we should exploit the available resources for effective transfer learning [244]. (ii) Compute intensity: it is environmentally unfriendly to domain-adaptive pretrain one model per language [215], and BioBERT was domain adaptive pretrained for 23 days on 8×Nvidia V100 GPUs. (iii) Ease of use: a single multilingual model eases deployment when an organization needs to work with multiple languages on a regular basis [97].

Our method, multilingual domain adaptive pretraining (mDAPT), extends domain adaptive pretraining to a multilingual scenario, with the goal of training a single multilingual model that performs, as close as possible, to N language-specific models. mDAPT starts with a base model, i.e. a pretrained multilingual language model, such as mBERT [59] or XLM-R [46]. As monolingual models have the advantage of language-specificity over multilingual models [189, 193], we consider monolingual models as upper baseline to our approach. We assume the availability of English-language domain-specific unlabelled text, and, where possible, multilingual domain-specific text. However, given that multilingual domain-specific text can be a limited resource, we look to Wikipedia for general-domain multilingual text [47]. The base model is domain adaptive pretrained on the combination of the domain-specific text, and general-domain multilingual text. Combining these data sources should prevent the base model from forgetting how to represent multiple languages while it adapts to the target domain.

Experiments in the domains of financial text and biomedical text, across seven languages: French, German, Spanish, Romanian, Portuguese, Danish, and English, and on two downstream tasks: named entity recognition, and sentence classification, show the effectiveness of multilingual domain adaptive pretraining. Further analysis in a cross-lingual biomedical sentence retrieval task indicates that mDAPT enables models to learn better domain-specific representations, and that these representations transfer across languages. Finally, we show

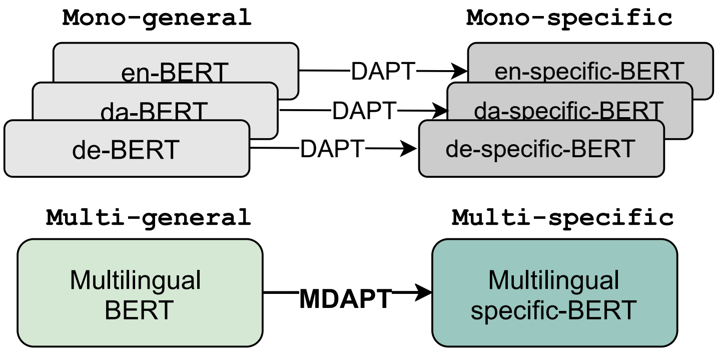


Figure 5: MDAPT extends domain adaptive pretraining to a multilingual scenario.

that the difference in tokenizer quality between mono- and multilingual models is more pronounced in domain-specific text, indicating a direction for future improvement.

All models trained with MDAPT and the new datasets used in downstream tasks and pretraining data[1] and our code is made available[2].

5.2 PROBLEM FORMULATION

Pretrained language models are trained from random initialization on a large corpus $\mathcal{C}$ of unlabelled sentences. Each sentence is used to optimize the parameters of the model using a pretraining objective, for example, masked language modelling, where, for a given sentence, 15% of the tokens are masked in the input m, and the model is trained to predict those tokens $J(\theta) = -\log p_\theta(x_m \mid \mathbf{x}_{\backslash m})$ [59]. $\mathcal{C}$ is usually a corpus of no specific domain,[3] e.g. Wikipedia or crawled web text.

Domain-adaptive pretraining is the process of continuing to pretrain a language model to suit a specific domain [83, 84]. This process also uses the masked language modelling pretraining objective, but the model is trained using a domain-specific corpus $\mathcal{S}$, e.g. biomedical text if the model should be suited to the biomedical domain. Our goal is to pretrain a *single model*, which will be used for downstream tasks in multiple languages within a specific domain, as opposed to having a separate model for each language. This single multilingual domain-specific model should, ideally, perform as well as language-specific domain-specific models in a domain-specific downstream task.

In pursuit of this goal, we use different types of corpora for domain adaptive pretraining of a single multilingual model. Each considered corpus has two properties: (1) a domain property – it is a `general` or

[1] https://github.com/RasmusKaer/mDAPT_supplements
[2] https://github.com/mahartmann/mdapt
[3] Text varies along different dimensions, e.g. topic or genre [183]. In the context of this paper, we focus on *domain-specificity* along the topic dimension, i.e. texts are considered as *domain-specific* if they talk about a narrow set of related concepts. The domain-specific text can comprise different genres of text (e.g. financial news articles and financial tweets would both be considered as being from the financial domain).

`specific` corpus; and (2) a language property – it is either `monolinugal` or `multilingual`. These properties can be combined, for example the multilingual Wikipedia is a `multi-general` corpus, while the abstracts of English biomedical publications would be a `mono-specific` corpus. Recall that `specific` corpora are not always available in languages other than English, but they are useful for adapting to the intended domain; while `multi-general` are more readily available, and should help maintain the multilingual abilities of the adapted language model. In the remainder of this paper, we will explore the benefits of domain adaptive pretraining with `mono-specific`, `multi-specific`, *and* `multi-general` corpora. Figure 5 shows how mDAPT extends domain adaptive pretraining to a multilingual scenario.

5.3 MULTILINGUAL DOMAIN ADAPTIVE PRETRAINING

Recall that we assume the availability of large scale English domain-specific and multilingual general unlabelled text. In addition to these `mono-specific` and `multi-general` corpora, we collect multilingual domain-specific corpora, using two specific domains—financial and biomedical—as an example (Section 5.3.1). Note that although we aim to collect domain-specific data in as many languages as possible, the collected data are usually still relatively small. We thus explore different strategies to combine different data sources (Section 5.3.2), resulting in three different types of pretraining corpora of around 10 million sentences, that exhibit `specific` and `multi` properties to different extents: **$\mathbf{E_D}$**: English domain-specific data; **$\mathbf{M_D}$+$\mathbf{E_D}$**: Multilingual domain-specific data, augmented with English domain-specific data; and **$\mathbf{M_D}$+$\mathbf{M_{WIKI}}$**: Multilingual domain-specific data, augmented with multilingual general data.

We use mBERT [59] as the multilingual base model, and employ two different continued pretraining methods (Section 5.3.3): adapter-based training and full model training, on these three pretraining corpora, respectively.

5.3.1 *Domain-specific corpus*

FINANCIAL DOMAIN As `specific` data for the financial domain, we use Reuters Corpora (RCV1, RCV2, TRC2),[4] SEC filings [58],[5] and FinMultiCorpus, which is an in-house collected corpus. The FinMultiCorpus consists of articles in multiple languages published on PwC website. The resulting corpus contains the following languages: *zh*, *da*, *nl*, *fr*, *de*, *it*, *ja*, *no*, *pt*, *ru*, *es*, *sv*, *en*, *tr*. Statistics on the presented languages can be found in Table 34 in the Appendix. Information about preprocessing are detailed in Appendix A.2.3.

BIOMEDICAL DOMAIN As `specific` data for the biomedical domain, we use biomedical publications from the PubMed database,

[4] Available by request at https://trec.nist.gov/data/reuters/reuters.html
[5] http://people.ischool.berkeley.edu/~khanna/fin10-K

Domain	Data	# Lang.	# Sent.	# Tokens
Fin	M_D	14	4.9M	34.4M
	E_D	1	10.0M	332.8M
	M_{WIKI}	14	5.1M	199.9M
Bio	M_D	8	3.2M	86.6M
	E_D	1	10.0M	370.6M
	M_{WIKI}	8	6.8M	214.2M

Table 12: A summary of pretraining data used. We use two specific domains—financial (top part) and biomedical (bottom part) as an example in this paper. M stands for Multilingual; E for English; D for Domain-specific; and, Wiki refers to general data, sampled from Wikipedia. The number of tokens are calculated using mBERT cased tokenizer. Note that because languages considered in financial and biomedical domains are not the same, we sample two different M_{WIKI} covering different languages.

in the following languages: *fr*, *en*, *de*, *it*, *es*, *ro*, *ru*, *pt*. For languages other than English, we use the language-specific PubMed abstracts published as training data by WMT, and additionally retrieve all language specific paper titles from the database.[6] For English, we only sample abstracts. We make sure that no translations of documents are included in the pretraining data. The final statistics on biomedical pretraining data can be found in Table 33 in the Appendix, as well as more details about preprocessing the documents. The descriptive statistics of these pretraining data can be found in Table 12.

5.3.2 *Combination of data sources*

Recall that `multi-specific` data is usually difficult to obtain, and we explore different strategies to account for this lack. The different compositions of pretraining data are illustrated in Figure 6. We control the size of the resulting corpora by setting a budget of 10 million sentences. This allows a fair comparison across data settings.

With plenty of English `specific` text available, E_D and M_D+E_D are composed by simply populating the corpus until reaching the allowance.

As a resource for `multi-general` data, we use Wikipedia page content, where we ensure the same page is not sampled twice across languages. Up-sampling M_D+M_{WIKI} using general domain multilingual data requires a sampling strategy that accounts for individual sizes. Sampling low-resource languages too often may lead to overfitting the repeated contents, whereas sampling high-resource language too much can lead to a model underfit. We balance the language samples using exponentially smoothed weighting [47, 59, 237]. Following Xue et al., we use a α of 0.3 to smooth the probability of sampling

[6] We use data from a bulk download of `ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline`, version 12/14/20

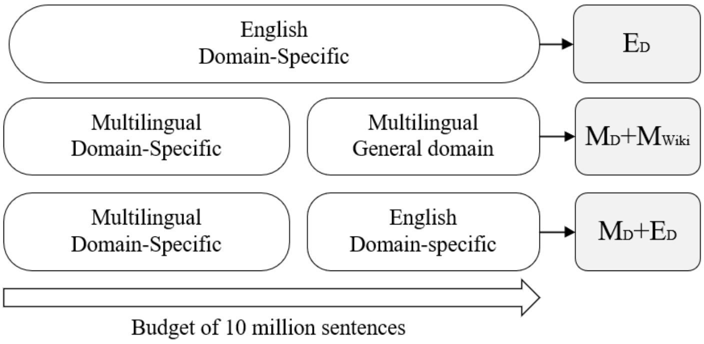

English Domain-Specific
ED
Multilingual Domain-Specific
Multilingual General domain
MD+MWiki
Multilingual Domain-Specific
English Domain-specific
MD+ED
Budget of 10 million sentences

Figure 6: Composition of pretraining data.

a language, $P(L)$, by $P(L)^{\alpha}$. After exponentiating each probability by α, we normalize and populate the pretraining corpus with Wikipedia sentences according to smoothed values until reaching our budget. Except for English, we up-sample using Wikipedia data. The statistics of the extracted sentences is presented in tables 33 and 34 in the Appendix.

5.3.3 *Pretraining methods*

CONTINUE PRETRAINING THE WHOLE MODEL We initialize our models with pretrained *base* model weights[7] and then continue pretraining the whole *base* model via the masked language modeling objective. We follow Devlin et al. [59] in randomly masking out 80% of subtokens and randomly replacing 10% of subtokens. For all models, we use an effective batch size of 2048 via gradient accumulation, a sequence length of 128, and a learning rate of 5e-5. We train all models for 25,000 steps, which takes 10 GPU days.

ADAPTER-BASED TRAINING In contrast to fine-tuning all weights of the *base* model, adapter-based training introduces a small network between each layer in the *base* model, while keeping the *base* model fixed. The resulting adapter weights, which can be optimized using self-supervised pretraining or later downstream supervised objectives, are usually much lighter than the *base* model, enabling parameter efficient transfer learning [89]. We train each adapter for 1.5M steps, taking only 2 GPU days. We refer readers to Pfeiffer et al. [167] for more details of adapter-based training and also describe them in the Appendix A.2.4 for self-containedness.

5.4 DOMAIN-SPECIFIC DOWNSTREAM TASKS

To demonstrate the effectiveness of our multilingual domain-specific models, we conduct experiments on two downstream tasks—Named Entity Recognition (NER) and sentence classification—using datasets from biomedical and financial domains, respectively.

[7] MBERT: https://huggingface.co/bert-base-multilingual-cased

		ncbi	**phar**	**quaero**	**clin**	**bioro**
		en	*es*	*fr*	*pt*	*ro*
# sents.	train	5,424	8,137	1,540	1,192	1,886
	dev	923	3,801	1,481	336	631
	test	940	3,982	1,413	973	629
# mentions	train	5,134	3,810	4,516	7,600	5,180
	dev	787	1,926	4,123	2,047	1,864
	test	960	1,876	4,086	6,315	1,768
# classes		1	4	10	13	4

Table 13: The descriptive statistics of the biomedical NER datasets.

5.4.1 *NER in the biomedical domain*

DATASETS We evaluate on 5 biomedical NER datasets in different languages. The French QUAERO [155] dataset, the Romanian BIORO dataset [145], and the English NCBI DISEASE dataset [61] comprise biomedical publications. The Spanish PHARMACONER [1] dataset comprises publicly available clinical case studies, and the Portuguese CLINPT dataset is the publicly available subset of the data collected by Lopes, Teixeira, and Gonçalo Oliveira [127], comprising texts about neurology from a clinical journal. The descriptive statistics of the NER datasets are listed in Table 13, and more details about the datasets can be found in Appendix A.2.2. We convert all NER annotations to BIO annotation format, and use official train/dev/test splits if available. For NCBI DISEASE, we use the data preprocessed by Lee et al. [120]. Further preprocessing details can be found in Appendix A.2.2.

NER MODEL Following Devlin et al. [59], we build a linear classifier on top of the BERT encoder outputs, i.e. the contextualized representations of the first sub-token within each token are taken as input to a token-level classifier to predict the token's tag. For full model fine-tuning, we train all models for a maximum of 100 epochs, stopping training early if no improvement on the development set is observed within 25 epochs. We optimize using AdamW, a batch size of 32, maximum sequence length of 128, and a learning rate of 2e-5. For adapter-based training, we train for 30 epochs using a learning rate of 1e-4.

5.4.2 *Sentence classification in the financial domain*

DATASETS We use three financial classification datasets, including the publicly available English FINANCIAL PHRASEBANK [138], German ONE MILLION POSTS [197], and a new Danish FINNEWS. The FINANCIAL PHRASEBANK is an English sentiment analysis dataset where sentences extracted from financial news and company press releases are annotated with three labels (Positive, Negative, and Neutral). Following its annotation guideline, we create FINNEWS—a dataset of Danish

	OMP	**FinNews**	**phr.bank**
	de	*da*	*en*
# sentences	10,276	5,134	4,845
# classes	2/9	3	3

Table 14: The descriptive statistics of the financial classification datasets. We frame the German dataset as a binary and a multi-class (9) classification tasks.

financial news headlines annotated with a sentiment. 2 annotators were screened to ensure sufficient domain and language background. The resulting dataset has a high inter-rater reliability (a measure of 82.1% percent agreement for raters and a Krippendorff's alpha of .725, measured on 800 randomly sampled examples). One Million Posts is sourced from an Austrian newspaper. We use title and topic for two classification settings on this dataset: a binary classification, determining whether a title concerns a financial topic or not; and a multi-class classification that classify a title into one of 9 topics. We list the descriptive statistics in Table 14, and further details can be found in Appendix A.2.3.

CLASSIFIER Following Devlin et al. [59], we built a classification layer on top of the [CLS] token. We perform simple hyperparameter tuning with the baseline monolingual model on each dataset separately. The parameter setting is selected on a coarse grid of batch-sizes $[16, 32]$ and epochs $[2, 4, 6]$. The best-performing hyperparameters on each dataset are then used in experiments using other pretrained models. All experiments follow an 80/20 split for train and testing with an equivalent split for model selection.

5.5 RESULTS

To measure the effectiveness of multilingual domain adaptive pretraining, we compare the effectiveness of our models trained with mDAPT on downstream NER and classification, to the respective monolingual baselines (`mono-general`), and to the base multilingual model without mDAPT (Table 15). Where available, we also compare to the respective monolingual domain-specific models (`mono-specific`).

BASELINE MODELS As `mono-general` baselines, we use English BERT [59], Portuguese BERT [212], Romanian BERT [64], BETO [34] for Spanish, FlauBert [117] for French, German BERT [36], and Danish BERT.[8] `Mono-specific` baselines exist only for a few languages and domains, we use EN-BIO-BERT [120] as English biomedical baseline, and EN-FIN-BERT [8] as English financial baseline. To the best of our knowledge, PT-BIO-BERT [199] is the only biomedical model for non-

[8] https://github.com/botxo/nordic_bert

	BIOMEDICAL NER					FINANCIAL SENTENCE CLASSIFICATION			
	quaero	**bioro**	**phar**	**ncbi**	**clinpt**	**omp-2**	**omp-9**	**finnews**	**phr.bank**
	fr	*ro*	*es*	*en*	*pt*	*de*	*de*	*da*	*en*
					FULL MODEL PRETRAINING				
MS-BERT	-	-	-	88.1	72.9	-	-	-	87.3
mono-BERT	**61.9**	**75.5**	88.2	85.1	72.6	**91.4**	71.5	**65.2**	**85.0**
MBERT	-3.7	-1.6	+0.2	+1.0	-0.2	-0.6	-0.4	-2.4	-2.6
+ E_D	-3.6	-1.6	**+0.6**	+1.5	-0.6	-0.3	0	-2.5	-1.2
+ M_D+E_D	-2.7	-0.9	+0.5	**+2.1**	**+0.1**	-0.2	**+0.1**	-1.6	-1.1
+ M_D+M_{WIKI}	-2.1	-1.4	+0.3	+1.8	0.0	-0.1	**+0.1**	-1.6	-1.4
					ADAPTER-BASED PRETRAINING				
mono-BERT	**58.6**	**73.2**	86.6	82.6	63.5	90.5	69.1	**66.0**	**85.3**
MBERT	-4.5	-4.5	-0.3	+0.1	-3.7	0.0	+0.8	-3.1	-3.1
+ E_D	-2.9	-2.0	+1.5	+1.4	+1.8	+0.7	+1.5	-4.9	-3.5
+ M_D+E_D	-1.3	-1.9	**+1.9**	+1.4	**+2.7**	**+0.9**	**+3.8**	-1.7	-2.6
+ M_D+M_{WIKI}	-1.4	-2.6	+1.0	**+1.8**	+1.6	+0.6	+2.6	-1.9	-3.2

Table 15: Evaluation results on biomedical NER and financial sentence classification tasks. We report the results—span-level micro F_1 for NER and sentence-level micro F_1 for classification—on the monolingual BERTs. Performance differences compared to the monolingual baselines are reported for multilingual BERTs, with and without MDAPT. All experiments are repeated five times using different random seeds, and mean values are reported. MS-BERT refers to mono-specific-BERT.

English language, we use it as Portuguese biomedical baseline, see Appendix A.2.1 for more details.

5.5.1 *Main results*

The main results for the biomedical NER and financial sentence classification tasks are presented in Table 15. We report the evaluation results for the mono-BERT baselines in the respective languages and the performance difference of the multilingual models compared to these monolingual baselines. We also consider two domain adaptive pretraining approaches: full model training, reported in the upper half of the table, and adapter-based training in the lower half.

Our work is motivated by the finding that domain adaptive pretraining enables models to better solve domain-specific tasks in monolingual scenarios. The first row in Table 15 shows our re-evaluation of the performance of the three available domain adaptive pretrained mono-specific-BERT models matching the domains investigated in our study. We confirm the findings of the original works, that the domain-specific models outperform their general domain mono-BERT counterparts. This underlines the importance of domain adaptation in order to best solve domain-specific task. The improvements of PT-BIO-BERT over PT-BERT are small, which coincides with the findings of Schneider et al. [199], and might be due to the fact that the CLINPT dataset comprises clinical entities rather than more general biomedical entities.

	mBERT	mDAPT	¬ mDAPT
quaero	58.2	59.8	58.0
bioro	73.9	74.5	73.4
ncbi	86.0	87.2	85.9
clin	72.4	72.7	71.8
phar	88.5	88.9	87.8
phr.bank	82.4	83.9	82.5
finnews	62.8	63.6	62.2
omp-2	90.8	91.3	91.0
omp-9	71.1	71.6	71.0/71.7

Table 16: Cross-domain control experiments. We report two control results for omp-9 since two mDAPT-setting achieved the same averaged accuracy.

full model training Recall that the aim of mDAPT is to train a single `multi-specific` model that performs comparable to the respective `mono-general` model. Using full model pretraining, we observe that the domain adaptive pretrained multilingual models can even outperform the monolingual baselines for *es* and *en* biomedical NER, and *de* for financial sentence classification. On the other hand, we observe losses of the multilingual models over the monolingual baselines for *fr* and *ro* NER, and *da* and *en* sentence classification. In all cases, mDAPT narrows the gap to monolingual performance compared to mBERT, i.e. multilingual domain adaptive pretraining helps to make the multilingual model better suited for the specific domain.

adapter-based training Adapter-based training exhibits a similar pattern: mDAPT improves mBERT across the board, except for the *da* and *en* sentence classification tasks, where mDAPT is conducted using only *en*-`specific` data. For most tasks, except *da* and *en* sentence classification, the performance of adapter-based training is below the one of full model training. On *pt* NER dataset, the best score (66.2) achieved by adapter-based training is much lower than the one (72.7) by the full model training.

comparison of combination strategies After we observe a single `multi` model can achieve competitive performance as several `mono` models, the next question is how do different combination strategies affect the effectiveness of mDAPT? As a general trend, the pretraining corpus composed of multilingual data—M$_{\mathrm{D}}$+E$_{\mathrm{D}}$ and M$_{\mathrm{D}}$+M$_{\mathrm{WIKI}}$—achieves better results than E$_{\mathrm{D}}$ composed by only *en* data. This is evident across both full - and adapter-based training. M$_{\mathrm{D}}$+E$_{\mathrm{D}}$ performs best in most cases, especially for the adapter-based training. This result indicates the importance of multilingual data in the pretraining corpus. It is worth noting that even pretraining only on E$_{\mathrm{D}}$ data can improve the performance on non-English datasets, and for *en* tasks, we see an expected advantage of having more *en*-`specific` data in the corpus.

5.5.2 *Cross-domain evaluations*

To make sure that the improvements of MDAPT models over MBERT stem from observing multilingual domain-specific data, and not from exposure to more data in general, we run cross-domain experiments [83], where we evaluate the models adapted to the biomedical domain on the financial downstream tasks, and vice versa. The results are shown in Table 16, where we report results for the best MDAPT model and its counterpart in the other domain (¬ MDAPT). In almost all cases, MDAPT outperforms ¬ MDAPT, indicating that adaptation to the domain, and not the exposure to additional multilingual data is responsible for MDAPT's improvement over MBERT. For the OMP datasets, ¬ MDAPT performs surprisingly well, and we speculate this might be because it requires less domain-specific language understanding to classify the newspaper titles.

5.6 ANALYSIS

Our experiments suggest that MDAPT results in a pretrained model which is better suited to solve domain-specific downstream tasks than MBERT, and that MDAPT narrows the gap to monolingual model performance. In this section, we present further analysis of these findings, in particular we investigate the quality of domain-specific representations learned by MDAPT models compared to MBERT, and the gap between mono- and multilingual model performance.

DOMAIN-SPECIFIC MULTILINGUAL REPRESENTATIONS Multilingual domain adaptive pretraining should result in improved representations of domain-specific text in multiple languages. We evaluate the models' ability to learn better sentence representations via a cross-lingual sentence retrieval task, where, given a sentence in a source language, the model is tasked to retrieve the corresponding translation in the target language. To obtain a sentence representation, we average over the encoder outputs for all subtokens in the sentence, and retrieve the k nearest neighbors based on cosine similarity. As no fine-tuning is needed to perform this task, it allows to directly evaluate encoder quality. We perform sentence retrieval on the parallel test sets of the WMT Biomedical Translation Shared Task 2020 [16]. The results in Table 17 show that MDAPT improves retrieval quality, presumably because the models learned better domain-specific representations across languages. Interestingly, with English as target language (upper half), the model trained on English domain-specific data works best, whereas for English as source language, it is important that the model has seen multilingual domain-specific data during pretraining.

	mBERT	$+E_D$	$+M_D+E_D$	$+M_D+M_W$
es → *en*	86.7	**91.9**	89.4	87.2
pt → *en*	**87.3**	77.1	77.5	83.9
de → *en*	79.4	**88.7**	83.9	80.9
it → *en*	85.6	**90.9**	87.4	87.1
ru → *en*	67.5	**84.4**	76.5	74.6
en → *es*	86.7	84.7	**90.5**	87.4
en → *pt*	89.4	78.2	**90.4**	86.8
en → *de*	79.4	79.6	**87.8**	81.2
en → *it*	83.9	82.9	**88.1**	86.1
en → *ru*	70.3	81.6	**90.8**	89.5

Table 17: Precision@1 for biomedical sentence retrieval. Best score in each row is marked in bold. The upper half shows alignment to English, the lower half alignment from English.

EFFECT OF TOKENIZATION Ideally, we want to have a mDAPT model that performs close to the corresponding monolingual model. However, for the full fine-tuning setup, the monolingual model outperforms the mDAPT models in most cases. Rust et al. [193] find that the superiority of monolingual over multilingual models can partly be attributed to better tokenizers of the monolingual models, and we hypothesize that this difference in tokenization is even more pronounced in domain-specific text. Following Rust et al. [193], we measure tokenizer quality via *continued words*, the fraction of words that the tokenizer splits into several subtokens, and compare the difference between monolingual and multilingual tokenizer quality on `specific` text (the train splits of the downstream tasks), with their difference on `general` text sampled from Wikipedia. Figure 7 shows that the gap between monolingual and multilingual tokenization quality is indeed larger in the `specific` texts (green bars) compared to the `general` texts (brown bars), indicating that in a specific domain, it is even harder for a multilingual model to outperform a monolingual model. This suggests that methods for explicitly adding representations of domain-specific words [173, 198] could be a promising direction for improving our approach.

ERROR ANALYSIS ON FINANCIAL SENTENCE CLASSIFICATION To provide a better insight into the difference between the `mono` and `multi` models, we compare the error predictions on the Danish FinNews dataset, since results in Table 15 show that the `mono` outperforms all `multi` models with a large margin on this dataset. We note that the FinNews dataset, which is sampled from tweets, contains a heavy use of idioms and jargon, on which the `multi` models usually fail. For example,

- Markedet lukker: **Medvind** til bankaktier på en rød C25-dag [Positive]

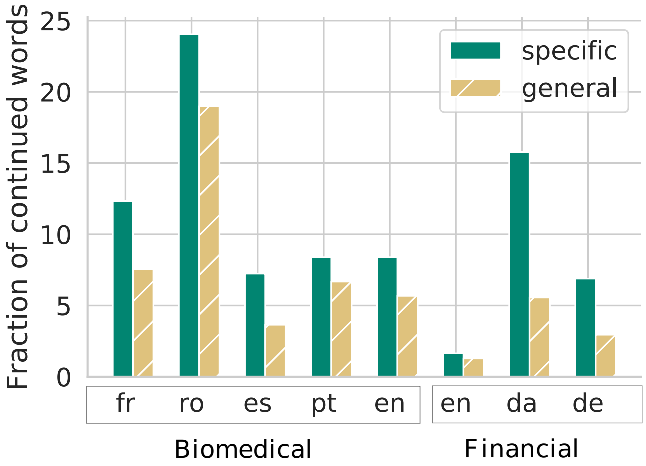


Figure 7: Difference in fraction of continued words between `mono`- and `multi`-lingual tokenizers on general and specific datasets. The bars indicate improvement of the monolingual tokenizer over the multilingual tokenizer.

English translation: *Market closes:* ***Tailwind*** *for bank shares on a red C25-day*

- Nationalbanken tror ikke særskat får den store betydning: Ekspert kaldet det **"noget pladder"** [NEGATIVE]

  English translation: *The Nationalbank does not think special tax will have the great significance: Expert called it "****some hogwash****"*

Pretraining data for the `mono` DA-BERT includes Common Crawl texts and custom scraped data from two large debate forums. We believe this exposes the DA-BERT to the particular use of informal register. By contrast, the pretraining data we use are mainly sampled from publications. This could be an interesting direction of covering the variety of a language in sub-domains for a strong MDAPT model.

5.7 RELATED WORK

Recent studies on domain-specific BERT [5, 120, 151], which mainly focus on English text, have demonstrated that in-domain pretraining data can improve the effectiveness of pretrained models on downstream tasks. These works continue pretraining the whole *base* model—BERT or RoBERTa—on domain-specific corpora, and the resulting models are supposed to capture both generic and domain-specific knowledge. By contrast, Beltagy, Lo, and Cohan [17], Gu et al. [82], and Shin et al. [203] train domain-specific models from scratch, tying an in-domain vocabulary. Despite its effectiveness, this approach requires much more compute than domain adaptive pretraining, which our work focuses on. Additionally, we explore an efficient variant of domain adaptive pretraining based on adapters [89, 167], and observe similar patterns regarding pretraining a multilingual domain-specific model.

Several efforts have trained large scale multilingual language representation models using parallel data [2, 47] or without any cross-lingual supervision [46, 59, 237]. However, poor performance on low-resource languages is often observed, and efforts are made to mitigate this problem [167, 174, 179]. In contrast, we focus on the scenario that the NLP model needs to process domain-specific text supporting a modest number of languages.

Alternative approaches aim at adapting a model to a specific target task within the domain directly, e.g. by an intermediate supervised fine-tuning step [168, 175], resulting in a model specialized for a single task. Domain adaptive pretraining, on the other hand, aims at providing a good base model for different tasks within the specific domain.

5.8 CONCLUSION

We extend domain adaptive pretraining to a multilingual scenario that aims to train a single multilingual model better suited for the specific domain. Evaluation results on datasets from biomedical and financial domains show that although multilingual models usually underperform their monolingual counterparts, domain adaptive pretraining can effectively narrow this gap. On seven out of nine datasets for document classification and NER, the model resulting from multilingual domain adaptive pretraining outperforms the baseline `multi-general` model, and on four it even outperforms the `mono-general` model. The encouraging results show the implication of deploying a single model which can process financial or biomedical documents in different languages, rather than building separate models for each individual language.

6 MULTIFIN: A DATASET FOR MULTILINGUAL FINANCIAL NLP

The following chapter is based on the article "MultiFin: A Dataset for Multilingual Financial NLP." by Rasmus Kær Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott, published in *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023 [99]. Appendix A.3 contains supplementary information for this chapter.

ABSTRACT

Financial information is generated and distributed across the world, resulting in a vast amount of domain-specific multilingual data. Multilingual models adapted to the financial domain would ease deployment when an organization needs to work with multiple languages on a regular basis. For the development and evaluation of such models, there is a need for multilingual financial language processing datasets. We describe MultiFin– a publicly available financial dataset consisting of real-world article headlines covering 15 languages across different writing systems and language families. The dataset consists of hierarchical label structure providing two classification tasks: multi-label and multi-class. We develop our annotation schema based on a real-world application and annotate our dataset using both 'label by native-speaker' and 'translate-then-label' approaches. The evaluation of several popular multilingual models, e.g., mBERT, XLM-R, and mT5, show that although decent accuracy can be achieved in high-resource languages, there is substantial room for improvement in low-resource languages.

6.1 INTRODUCTION

Natural language processing technology has substantially improved in recent years due to the general-purpose Transformer model [224], large-scale self-supervised training from unlabelled corpora [59], and the scaling of both of these to increasingly large datasets and models [178]. Nevertheless, there are still benefits to having domain-specific models [83], especially when working with clinical [52] or financial text [8].

The domain of financial text is particularly interesting for multilingual NLP, given that it is produced across the world [101, 121]. The text often includes invoices, transactions, accounting data, tax policies, and stock market information, *inter-alia*, and there is an emerging effort to create monolingual financial BERTs (FinBERTs) to pro-

Example	Lang.	Low-Level	High-Level
Encuesta Mundial de CEOs 2019 - Hostelería	SPA	· BOARD · RETAIL	Business & Management
Amendments to VAT legislation	ENG	· VAT · GOV	Tax & Accounting
Skatta- og lögfræðisvið	ISL	· TAX	Tax & Accounting
Bestyrelsens rolle i forhold til strategiarbejdet	DAN	· BOARD	Business & Management
Εισαγωγη στην Ελληνικη Φορολογία	GRE	· TAX	Tax & Accounting
「事業再編・再生支援」と「ディール戦略」部門を統合・強化	JPN	· M&A · BOARD	Finance
Veri Analitiği ve Adli Bilişim Çözümleri	TUR	· FINCRIME · TECH	Government & Controls

Table 18: Examples from the MultiFin dataset covering different languages, writing scripts, and combinations of Low-Level and High-Level labels. See Section 6.3 for more details on the languages and annotation process. Abbreviations for Low-Level labels: Board, Strategy & Management (BOARD), M&A & Valuations (M&A), Financial Crime (FINCRIME), Technology (TECH), Government & Policy (GOV), Retail & Consumers (RETAIL), and VAT & Customs (VAT).

cess financial text [8, 54, 126, 239]. However, the handling of financial text by multinational companies is inherently multilingual, therefore, there is is a need for datasets to evaluate how well models can process multilingual financial text.

To this end, we introduce the MultiFin dataset, a publicly available financial dataset consisting of real-world financial article headlines in 15 languages (see examples in Table 18). MultiFin is annotated with High-Level and Low-Level topics for multi-class and multi-label classification, respectively. The dataset is intended as a resource for developing multilingual financial language models. It is the first benchmark for evaluating cross-lingual and multilingual performance of financial models across multiple languages, writing systems and language families that reflects the real-world multilingual situation in the financial domain.

We benchmark four large-scale pretrained language models (SentenceBERT, mBERT, XLM-R, and MT5) and find that the benefits of large-scale pretraining also apply to financial text. XLM-R is clearly the best performing model in all of our experiments, however, there is a subsantial gap in performance between high- and low-resource languages in MultiFin. Moreover, a simple LSTM initialized with FastText word embeddings gives surprisingly competitive performance in several experiments. Overall, we find the financial domain can benefit from multilingual NLP, and future work should focus on domain adaptive efforts and improving models' capacity to generalize to low-resource languages.

CONTRIBUTIONS Our contributions are as follows: (a) We present a multilingual financial dataset based on article titles in multiple languages and annotated with two levels of topics. The dataset is made publicly available at https://github.com/RasmusKaer/MultiFin. (b) We evaluate different multilingual models under different setups in conjunction with analysis on the multilingual MultiFin to establish baselines for the benchmark. (c) Our analysis identifies a need for further research in minimizing the performance gap between high and low-resource languages, and domain adaptive efforts maybe be a promising direction for narrowing this gap.

6.2 EXISTING DATASETS FOR FINANCIAL NLP

Financial NLP is an emerging area of NLP. Researchers and practitioners have a keen interest in processing natural language for different downstream tasks in the financial domain, such as text mining in accounting [129], financial transactions [102], sentiment analysis [138], and text classification [9]. Also, financial economics research shows that news articles and media can be used to forecast firm performance [218], predict stock market volatility [77] and predict market return [217]. Moreover, Qin and Yang [176] show that textual transcripts in combination with audio recordings of company earnings conference calls can be used to predict stock price volatility.

There is a large variety of downstream NLP tasks in the financial domain. However, most work within the community is carried out in a monolingual English setting, where the focus is on adapting successful generic monolingual models to the financial domain [8, 54, 126, 239]. Only a little work on multilingual domain-adapted models has been investigated [101]. Since the financial environment is indeed multilingual, further progression is conditioned on the availability of multilingual resources to develop new methods for multilingual NLP in the financial domain.

DATASETS IN THE FINANCIAL DOMAIN An extensive literature review identifies the datasets used for financial NLP. We define three criteria for being assigned to the list: (1) the dataset needs to be publicly available and accessible, (2) it needs a clear definition of the task with accompanying annotations (i.e., labels, tags, etc.), and (3) it needs to be peer-reviewed and documented. These criteria are set to ensure the quality of the data resource and proper availability and accessibility. Table 19 presents our findings.

An investigation of the datasets shows that most resources are in English. Table 19 (A) presents an overview of the English evaluation datasets. AnalystTone Dataset [92], FinTextSen [50] and Financial Phrase Bank [138] are among the most popular datasets. Sentiment analysis is the most frequent task for the datasets, followed by

(A) Datasets in English		**(B) Non-English datasets**		lang
AnalystTone Dataset [92]	SA	DanFinNews [101]	SA	dan
FinTextSen [50]	SA	CorpusFR [95]	NER,RE	fre
Financial Phrase Bank [138]	SA	BORSAH [6]	SA	ara
FiQA Dataset [136]	SA,QA			
FinNum-1 [38]	Numeral CLS	**(C) Multilingual datasets**		
M&A dataset [238]	Deal completeness CLS	eng-chi Parallel Fin. Dataset [223]	TC,MT	eng,chi
FinNum-2 [37]	Numeral attachment	FNS-2022* Shared Task [67]	SA	eng,spa,gre
StockSen* [235]	SA	SEDAR* [75]	MT	eng,fre
FinCausal* [139]	RC,RE	FinSBD-2019* [13]	SBD	eng,fre
MultiLing2019 [66]	Summarization	SIXX-Corpora* [73]	SA	eng,spa,ger
FIN5 & FIN3 [195]	NER			
Stock-event [119]	Stock Price Prediction	**(D) Our dataset**		
News-sample OMX Helsinki* [137]	SA	MultiFin (this paper)	TC	eng,dan,fin,gre,heb,hun,isl,
EarningsCall [176]	Stock Price Volatility			ita,jpn,nor,pol,rus,spa,swe,tur
Stocknet [236]	Stock Movement Prediction			

Table 19: A list of datasets for financial NLP with corresponding task (SA=Sentiment Analysis, NER=Named Entity Recognition, QA=Question Answering, TC=Topic Classification, RC=Relation Classification, RE=Relation Extraction, MT=Machine Translation, SBD=Sentence Boundary Detection, CLS=Classification). Marked (*) refers to datasets where a request is needed or an application for permission needs to be obtained before that dataset is shared.

classification. Only few non-English and multilingual datasets exist. Table 19 (B) and (C) shows available datasets in other languages than English. There are five multilingual datasets which contain English plus three additional non-English languages. The dataset containing most languages is the trilingual datasets FNS-2022 Shared Task [67] and SIXX-Corpora [73]. In addition, we found three low-resource monolingual sentiment datasets: Arabic BORSAH [6], Greek FNS-2022 Shared Task [67] and the Danish DanFinNews [101] which is the Danish equivalent to the Financial PhraseBank.

The need for a multilingual financial resource has been highlighted in several studies [73, 95, 101] and its lack of multilingual resources is a limitation for further progression. There is also a need for including different language families and low-resources languages into the research landscape to ensure that not only the high-resources languages lays the foundation of research [6]. This suggests a gap in resources necessary to advance the financial NLP towards a more multilingual scenario that simulate the financial domain's multilingual environment. Our work, see Table 19 (D), is motivated by creating a gold standard for benchmarking financial models to facilitate work on adapting to multiple languages within a specific domain.

6.3 THE MULTIFIN DATASET

The MultiFin dataset is a multilingual corpus, consisting of real-world article headlines covering 15 languages. We annotate the corpus using hierarchical label structure, providing two classification tasks: multi-class and multi-label classification.

DATA COLLECTION The dataset builds on a collection of public articles published on a large accounting firm's websites. A subset of

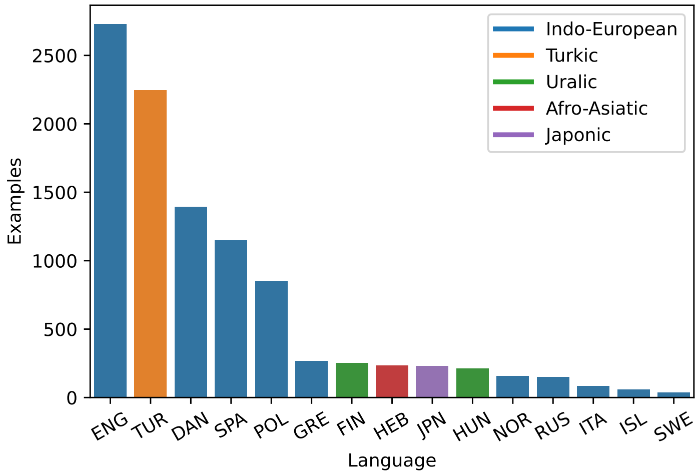


Figure 8: Number of examples per language in MultiFin. Bars in the same color indicate these languages belong to the same language family. In this paper, we define languages with more than 500 examples—ENG, TUR, DAN, SPA, POL—high resource languages and the remaining low resource languages.

the archive was made available for this study. The data collection is based on a real-world application deployed in a large accounting firm. The language selection is determined by the company branches that made their data available to us. We build a multilingual dataset from the headlines of the entire subset that the firm made available. The subset of the archive covers published material in 15 languages and comprises around 10K headlines. The distribution of headlines over languages is shown in Figure 8. The publication date is mainly from the period of 2015 to 2021 with some titles having missing dates. The proposed benchmark contains all the languages we were permitted to use, reviewed by experts, which ensures the reliability and quality of both language and content. While the selection of the 15 languages might not be ideal (e.g., African and Indic languages as well as Arabic and Modern Standard Mandarin are missing), we provide the first massively multilingual dataset for financial NLP, see Table 19 for an overview over currently available datasets. It is also worthy noting that headlines, due to their limited context, poses a great challenge for text classification models deployed in the wild [41]. See Figure 13 for the text length distribution across different languages.

ANNOTATION SCHEME The articles were already tagged with internally pre-defined topics from a company-internal system. Based on these topics, we derive a new, more general label set, referred to Low-Level. Through our label scheme we seek to have different levels of granularity since it gives us the opportunity to go deeper into evaluating the ability of identifying the more refined topics that are presented in titles. Therefore, we first assign fine-grained tags to the topics contain in an headline. For this we use the Low-Level topics. Secondly, we also assign the headline to a single more coarse-grained category, referred to High-Level. We defined the High-Level topics on the basis of universal categories typically found in news media and more content categorization. Our fine-grained annotation process results in a dataset with multiple labels per headline. We derive HIGH-LEVEL single labels from these multi-label annotations based on either a majority-vote, using the first tag in case of ties. The overview of LOW-LEVEL and HIGH-LEVEL topics is presented in 20.

HIGH-LEVEL	LOW-LEVEL
Technology	Technology
	IT Security
Industry	Power, Energy & Renewables
	Supply Chain & Transport
	Healthcare & Pharmaceuticals
	Retail & Consumers
	Real Estate & Construction
	Media & Entertainment
Tax & Accounting	VAT & Customs
	Tax
	Accounting & Assurance
Finance	M&A & Valuations
	Asset & Wealth Management
	Actuary, Pension & Insurance
	Banking & Financial Markets
Government & Controls	Government & Policy
	Financial Crime
	Governance, Controls & Compliance
Business & Management	Board, Strategy & Management
	Start-Up, Innovation & Entrepreneurship
	Corporate Responsibility
	SME & Family Business
	Human Resources

Table 20: Overview of HIGH-LEVEL and LOW-LEVEL topics. The coarse-grained single labels are derived from the fine-grained multi-label annotations based on either a majority-vote, using the first tag in case of ties.

ANNOTATION PROCESS We ask native-level speakers of English and Danish to annotate the dataset using the LOW-LEVEL tags. The annotators have domain expertise and participated on a voluntary basis. Detailed annotation guidelines were presented to the annotators before they started. The description contains definitions of topics including some exemplifications of themes and concepts that may occurs for the topics. As for the annotation of multiple labels, the annotators were asked to label up to three topics per example. The annotated labels needed to be ordered by topic weight, i.e., the first annotated topic is the most dominating topic in the sentence, then the second and third most. The overview and statistics of the label distributions can be found in Appendix A.3.2.

TRANSLATE-THEN-LABEL EVALUATION We translated the headlines into English for topic annotation using a translation service.[1] We carefully assessed the translation quality to ensure that the translation process does not introduce noise into our dataset. We want to check whether the content of the original sentence is contained in the translation to English. That is, the topics or matters treated in an article stay the same for the translation. For the evaluation, we randomly sample 50 examples from DAN, NOR, ITA, SPA, POL and the entire SWE. We asked evaluators with language proficiency to assess the samples. We presented them with the original sentence, its English translation, and the annotated topics, and ask to answer a true/false question of 1) is the content of the original sentence contained in the English translation, 2) is the property that makes the English sentence fall into this category present in the original sentence as well?

The evaluation shows that for DAN, NOR, ITA, SPA, POL and SWE all preserved the properties that make the article fall into a specific category. There was not reported any errors by the evaluators. Thus, we consider translation quality to be high enough to not introduce noise in the process.

ANNOTATOR AGREEMENT Inter-annotator agreement is measured as multi-label Cohen's κ [45]. The sample selected for evaluation by both annotators is 1200 examples, randomly sampled across languages and topics. The combined κ of 0.94 suggests a a near-prefect agreement. Table 35 depicts the topic-level κ.

DESCRIPTION OF DATASET The dataset consists of 10,048 headlines in 15 languages annotated with 23 topic labels for LOW-LEVEL and 6 HIGH-LEVEL topics for multi-class. See Appendix A.3.2 for details on the distribution of the LOW-LEVEL topics and HIGH-LEVEL topics and Appendix A.3.6 for an overview of the sentence length distribution across different languages. For multi-class, multi-label classification, we have a total of 14,230 tags across 10,048 headlines (80,678 tokens) using 23 fine-grained topics. For multi-class, single label, we have a coarse-grained topic tag for each headline.

6.4 EXPERIMENTS AND RESULTS

We employ popular pre-trained multilingual models[2] and test their effectiveness under different experimental setups. For experimentation, we will only focus on the LOW-LEVEL multi-label task, and HIGH-LEVEL results are reported in the Appendix, Table 39.

[1] Google Translate, version as of Autumn 2021.

[2] The number of trainable parameters for each model is listed in Table 38 in the Appendix.

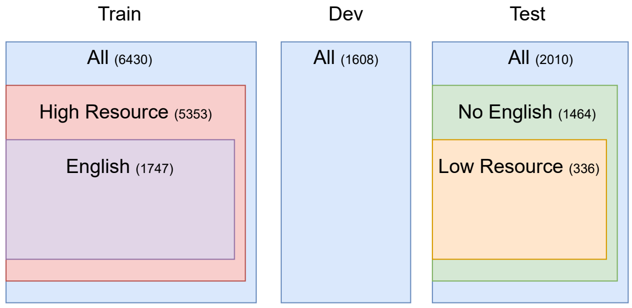


Figure 9: We train models on the complete training set as well as two subsets, to evaluate the multilingual learning and cross-lingual transfer capacities, respectively. We use a joint development set of all the languages to select the trained checkpoint. The final model is evaluated on the test and metrics evaluated on the complete test as well as two subsets are reported. Numbers in brackets are the examples belonging to the corresponding (sub)set.

6.4.1 *Models*

mBERT [59] has been pre-trained on Wikipedia articles of 104 languages. Similarly, **XLM-R** [46] was pre-trained on web crawl data, whose size is much larger than Wikipedia data. For both mBERT and XLM-R, we built a classification layer on top of sentence embedding (i.e., the hidden states corresponding to the first [CLS] token). The classification layer consists of a dense layer and tanh activation function, followed by another dense layer, where the output dimension is the total number of possible topics.

sBERT [186] we use multilingual sentence BERT to map an input sentence to a 768 dimensional dense vector space and then build a classification layer on top of it. Note that we follow Reimers and Gurevych [185] to keep the weights of sBERT fixed and use sBERT as a feature extractor. We also investigate the variant of fine-tuning sBERT together with the classification layer. The results of fine-tuning approach are very close to feature extraction approach, although the latter involves much smaller number of trainable parameters (110M vs 600K).

mT5 [237] was pre-trained on web crawl data covering 101 languages using a 'text-to-text' format. That is, consecutive spans of input tokens are replaced with a mask token, and then an encoder-decoder transformer is trained to reconstruct the masked-out tokens. When MT5 is used for down-stream classification task, the model outputs the literal text of the label instead of a class index. In addition to these transformer-based models, we also experiment with models using pre-trained type-based embeddings described below.

ALIGNED FASTTEXT EMBEDDINGS As a baseline, we experiment with models using pre-trained type-based embeddings[3], in particular

[3] Fasttext models enable the computation of embeddings for out-of-vocabulary words based on sub-tokens.

Model	Training	Test		
		ALL	NO ENGLISH	LOW RESOURCE
FASTTEXT$_{\text{BAG}}$	ALL	74.2 ± 0.2	71.7 ± 0.2	60.9 ±0.8
	ENGLISH	41.8 ± 1.5	24.5 ± 1.6	27.9 ± 3.2
	HIGH RESOURCE	70.3 ± 1.1	66.8 ±1.1	38.2 ± 1.2
FASTTEXT$_{\text{LSTM}}$	ALL	85.4 ± 0.4	83.6 ± 0.4	74.4 ± 0.9
	ENGLISH	51.6 ± 0.5	36.9 ± 0.6	41.9 ± 1.9
	HIGH RESOURCE	82.4 ± 0.6	80.0 ± 0.6	59.5 ± 1.5
SBERT	ALL	73.5 ± 0.2	67.9 ± 0.2	52.0 ± 0.2
	ENGLISH	50.8 ± 0.5	32.7 ± 0.4	27.5 ± 0.6
	HIGH RESOURCE	69.9 ± 0.3	62.8 ± 0.5	27.4 ± 0.2
MBERT	ALL	88.6 ± 0.3	86.5 ± 0.3	77.9 ± 0.5
	ENGLISH	58.3 ± 0.7	43.5 ± 1.0	39.4 ± 2.3
	HIGH RESOURCE	84.1 ± 0.4	80.6 ± 0.4	47.7 ± 0.7
XLM-R	ALL	**90.8** ± **0.4**	**89.4** ± **0.4**	**83.9** ± **0.6**
	ENGLISH	68.0 ± 1.3	59.2 ± 1.6	59.8 ± 1.9
	HIGH RESOURCE	88.6 ± 0.4	86.4 ± 0.5	71.0 ± 1.9
MT5	ALL	81.3 ± 0.1	76.6 ± 0.2	51.0 ± 1.5
	ENGLISH	50.7 ± 1.0	34.3 ± 1.1	25.5 ± 1.9
	HIGH RESOURCE	78.5 ± 0.3	72.9 ± 0.5	33.7 ± 0.2

Table 21: Evaluation results on fine-grained topics (LOW-LEVEL). This is a multi-label classification task with 23 labels, and each example may be assigned up to three topics. All experiments are repeated five times using different random seeds. Averaged Micro F_1 scores and the standard deviations are reported. Best results per column are marked in bold.

the 300-dimensional fasttext embeddings [22] trained on Common Crawl and Wikipedia data [81]. In order to enable cross-lingual transfer, we map language-specific fasttext embeddings for all languages covered in our dataset into a space[4], using RCSLS [103] as a supervised mapping method. Details about embedding alignment can be found in Appendix A.3.3. The mapped embeddings are used as inputs for two baseline models: an LSTM classifier (FASTTEXT$_{\text{LSTM}}$) and a bag-of-embeddings (FASTTEXT$_{\text{BAG}}$) classifier. The LSTM classifier consists of one bidirectional LSTM layers with a classification layer on top, which receives as input a concatenation of the final hidden states of the top-most layer of forward and backward LSTM. The BoE classifier uses the average over all word embeddings in the input sequence as input to the classification layer. For both models, we use the same classification layer as for the MBERT and XLM-R models.

[4] We compute pairwise mappings between non-English source embeddings and English target embeddings, and map all non-English embeddings into the space of English embeddings.

6.4.2 *Experimental setup*

To evaluate multilingual learning, we train the model on the complete training set that contains all 15 languages (referred to as All). To evaluate cross-lingual transfer, we train the model on (i) a subset that contains only English training data (English); and, (ii) a subset that contains 5 high-resource languages (i.e., English, Turkish, Danish, Spanish, Poland) (High Resource).

MODEL SELECTION In the context of zero-shot cross-lingual transfer, it was shown that performance on a source language (e.g., English) development set does not correlate well with performance in the target language [42, 108]. We follow Conneau et al. [49] and use a joint development set of all the languages. Figure 9 is a high-level illustration of our experimental setup. The trained model which achieves the highest Micro F_1 score on the development set is finally evaluated on the test set. We repeat all experiments five times using different random seeds and mean values and standard deviations are reported.

6.4.3 *Results*

Table 21 shows that models trained on the training set consisting of all languages (All) achieve slightly better results (2.0-4.5 absolute F_1) than the ones trained on high-resource languages (High Resource) when the trained models are evaluated on the complete test set. However, this performance gap becomes much larger (11.4-30.2 absolute F_1) when models are evaluated on the subset containing only low-resource languages, which is expected, as the latter setting requires zero-shot transfer when training on High Resource and evaluating on Low Resource.

In the per language analysis (detailed in the following section), we also observe that once the training set contains abundant examples (500+) for these languages, models achieve nearly the same results when evaluated on high-resource languages (Figure 10). Therefore, we focus our discussion on the evaluation results on low-resource languages.

The first observation is that different pre-trained multilingual models differ in multilingual learning abilities on our dataset. That is, when they are fine-tuned on All, model effectiveness on low-resource languages ranges from 51.0 to 83.9 (A detailed analysis can be found in the following section).

The ability of zero-shot cross-lingual transfer is another interesting property of multilingual models. Previous studies show that models trained on English only can achieve impressive results on examples in other languages [49, 91]. However, we observe poor performance when models are trained on English and evaluated on Low Resource (all under 40 F_1 except XLM-R achieving near 40 F_1). In terms of the choice of source languages, we observe moderate improvements (6.8-11.2 F_1) when massively multilingual pre-trained models (i.e., MBERT, XLM-R, MT5) are cross-lingual transferred from more languages (HIGH RESOURCE: ENG, TUR, DAN, SPA, POL) rather than from ENGLISH only. On the other hand, the improvement becomes much larger (17.6 F_1) when FASTTEXT$_{\text{LSTM}}$ is trained on more languages, indicating that the model might make better use of information from additional languages than the transformer-based models. When training on HIGH RESOURCE, FASTTEXT$_{\text{LSTM}}$ only slightly underperforms MBERT, and outperforms all other models except XLM-R for transfer from HIGH RESOURCE to LOW RESOURCE. This might be due to the explicit embedding alignment mechanism used in the FASTTEXT approach.

We also calculated the Wilcoxon signed-rank test to assess whether there is a statistically significant difference between the results of XLM-R and MBERT. XLM-R significantly (p-value $\leqslant 0.05$) outperformed MBERT when trained on ALL, ENGLISH, and HIGH RESOURCE and then evaluated on the complete test set. However, the differences for individual languages were not always statistically significant ($p > 0.05$). When both models were trained on ALL, the differences in performances on TUR, NOR, RUS, SWE, ITA, and ISL were not significant; the same holds for the difference on ENG when trained on ENGLISH as well as for the differences on SWE and ISL when trained on HIGH RESOURCE.

6.5 ANALYSIS AND DISCUSSION

Our experiments suggest that although decent accuracy can be achieved for high-resource languages, there is substantial room for improvement in achieving better performance on the multilingual financial dataset. In this section, we present a detailed analysis of the results and investigate some of the findings to identify possible modelling improvements and look into the different dimensions of our dataset.

6.5.1 *Multilingual abilities from a language-level perspective*

Multilingual models should ideally learn good representations for all languages they were pre-trained on but this is difficult to achieve in practice due to the "curse of multilinguality" [46]. Figure 10 presents per-language results for the three training settings ALL, ENGLISH, and HIGH RESOURCE. Generally, we see that XLM-R outperforms the rest of the models across all test settings and languages. When training on ALL data (first block in Figure 10), although the models have seen all languages during training, MT5 and sBERT seem to be struggling particularly with GRE, JPN, HEB and HUN. We see a drop in performance between high (upper part of the column) and low-resource languages (bottom part of the column), which is expected as the low-resource languages have less examples in the training dataset. When training on HIGH RESOURCE (last block in Figure 10), we observe that performance for the high-resource languages seen during training is

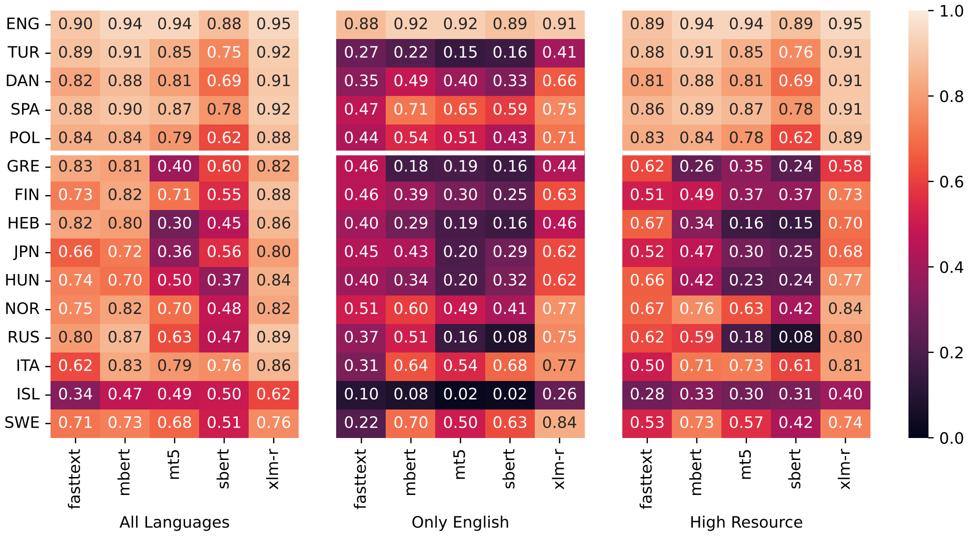


Figure 10: Per language analysis with the multi-label, LOW-LEVEL setting. We train on the three settings: ALL, ENGLISH, and HIGH RESOURCE and test on ALL. The first column in each block refers to the FASTTEXT$_{\text{LSTM}}$. Languages are in descending order by the number of examples in MULTIFIN, with a white separator between high and low-resource languages.

stable compared to training on ALL (indicating that including low-resource languages during fine-tuning does not hurt performance on high-resource languages), but performance for zero-shot transfer to low-resource languages drops significantly. We compare the performance drops suffered on low resource languages from training on ALL data to training on HIGH RESOURCE data between XLM-R, MBERT, and FASTTEXT$_{\text{LSTM}}$, and find that MBERT suffers from larger performance drops than the other models for most languages, with the largest drops for GRE and HEB. XLM-R shows the smallest performance drops for most languages, indicating that it has better zero-shot transfer abilities than the other models.

Next, we analyze the best source for zero-shot transfer by comparing the performance on low-resource languages for models trained on HIGH RESOURCE data with models trained on ENGLISH data. In all cases (except XLM-R on SWE), zero-shot transfer works better when more languages are included in the training set. This might be due to the fact that training on more languages allows models to learn more robust representations of input sequences. Another factor might be that, as our dataset has a large label space, including more training examples (regardless of language) can improve learning representations of otherwise sparse classes. As indicated by the averaged results reported in the previous section, for most languages (except FIN and ISL), FASTTEXT$_{\text{LSTM}}$ shows higher improvements when including more languages to train on.

Comparing zero-shot performance on different target languages for models trained on ENGLISH (middle block in Figure 10) reveals that all models with a slight exception to XLM-R struggle to generalize to languages not seen during fine-tuning, although they were part of the pre-training languages. Previous research on mBERT suggests a correlation between zero-shot performance in a downstream task and amount of in-language pre-training data [115, 232], which we also observe in our results. Overall, we see very poor generalization ability to certain low-resource languages, such as ISL, GRE, HEB, and RUS. Particularly for ISL, transfer ability from English is nearly non-existing, indicating a need for multilingual models with better transfer abilities to low-resource languages.

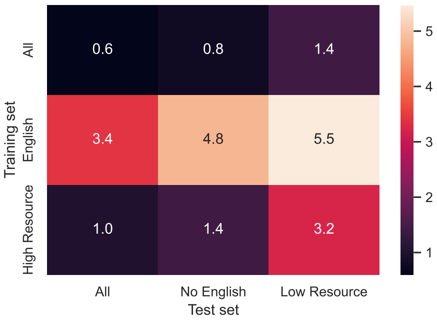


Figure 11: The improvement over the vanilla mBERT, in Micro F1, due to domain-adaptive pre-training mBERT. We compare the model by Jørgensen et al. [101] against the vanilla mBERT.

6.5.2 *Domain-adaptive pre-training can boost the cross-lingual performance*

Domain-adaptive pre-training has been shown to improve the model effectiveness when these models are employed to process domain-specific text [83]. We evaluate the publicly available model by Jørgensen et al. [101], which continues pre-training mBERT on the combination of multilingual financial text and Wikipedia, and measure the improvement over the vanilla mBERT in Table 21. Note that the multilingual pre-training data in [101] cover 9 languages in MultiFin, except POL, GRE, FIN, HEB, HUN, and ISL. Nevertheless, results in Figure 11 shows that domain-adaptive pre-trained models outperform vanilla mBERT in all experimental setups, and larger improvements are observed when training set and test set are disjoint, for example, when models are trained on English or high-resource languages and tested on low-resource languages.

6.5.3 *Multilingual versus translate*

We assessed that the translation quality was good enough to preserve the topics in Section 6.3. Therefore, we translate all training and test data to English and fine-tune a monolingual model for English (RoBERTa, Liu et al. [124]) on the translated training data. We compare performance on the translated test sets with XLM-R trained and tested on the multilingual data.

The monolingual model's advantage of language-specificity over multilingual models [189, 193] is evident in Figure 12, where the

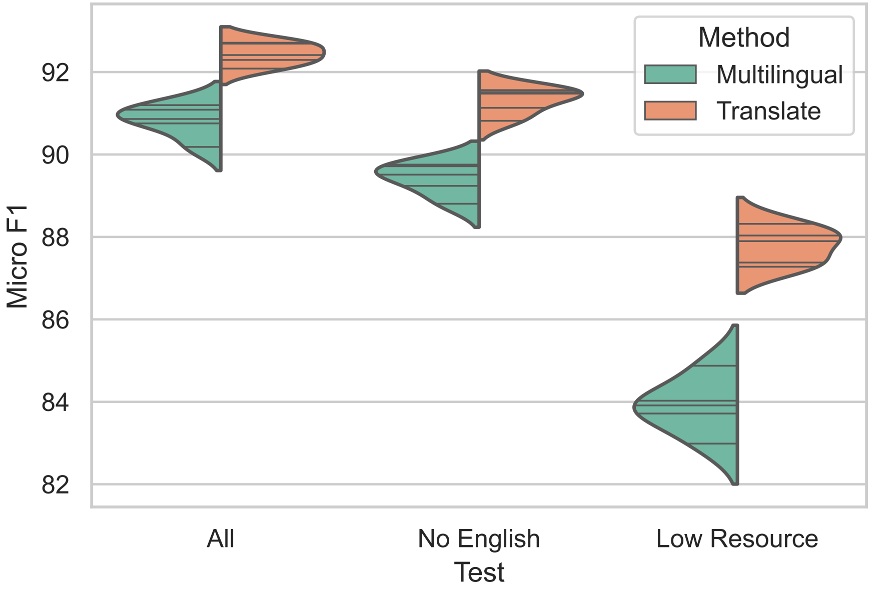


Figure 12: Multilingual (i.e., XLM-R) against translate approach based on English RoBERTa. We use the same setting as in Table 21, where we train on all languages and test on All Lang., NoEnglish and LowRes.

monolingual model trained on English is slightly better than the multilingual model trained on multilingual data.[5] We consider this monolingual model an additional baseline on MultiFin.

6.6 CONCLUSION

We proposed MultiFin, a dataset for the evaluation of multilingual financial NLP models. The main aim is to advance multilingual NLP in the financial domain so it is better suited for new development and evaluation of domain-specific models. MultiFin is a diverse dataset with 10,000 examples, covering 15 languages, including different language families and writing systems. We benchmark a collection of standard multilingual language models on MultiFin and find that although these models often achieve good performance in high-resource languages, there is a substantial gap in performance between high- and lower-resource languages. The per-language analysis uncovered that most of the benchmarked models do not facilitate a good transfer across the evaluated languages, and for specific languages, indicate a strong need for improving the models' capacity to generalize. The multilingual mDAPT model presented overall better generalization, particularly to low-resource languages, indicating that focusing on multilingual domain-specific methods is a promising direction for

[5] Artetxe, Labaka, and Agirre [10] found that improvements of a translation baseline in a cross-lingual NLI task do not stem from overcoming the cross-lingual gap, but from the fact that translation of the training data introduces alterations which improve generalization to a translated test set. It is possible that in our experiments, the performance of the monolingual model generalizing from translated training data to translated test data is impacted by similar mechanisms.

future work in financial NLP. Future work includes extending the dataset to include more examples across more languages so better understand the limits of multilingual financial text processing. We are also exploring including the entire document, as opposed to only the headline, but this would depend on high-quality long document processing models [52]. We hope to motivate and inspire collective work on multilingual NLP in the financial domain.

limitations

annotators We are aware that annotators with domain knowledge and language proficiency would be preferred. It was not within our resources to find qualified annotators in the financial domain with expert knowledge and language proficiency for all 15 languages.

annotation process The number of annotated topics per example is determined to three, although a handful of article titles could potentially be assigned more than three topics. The authors attempted to limit this by prioritizing annotated topics by topic weight (see Section 6.3).

acknowledgements

We thank PwC for providing the data and thank Lars Silberg Hansen for his support and valuable contribution to the creation of this dataset.

Part IV

EXPLAINABILITY IN MULTILINGUAL NLP

7 ARE MULTILINGUAL SENTIMENT MODELS EQUALLY RIGHT FOR THE RIGHT REASONS?

The following chapter is based on the article "Are Multilingual Sentiment Models Equally Right for the Right Reasons?" by Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard, published in *Proceedings of the Fifth BlackBoxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 131–141 [100][1]. Appendix A.4 contains supplementary information for this chapter.

ABSTRACT

Multilingual NLP models provide potential solutions to *the digital language divide*, i.e., cross-language performance disparities. Early analyses of such models have indicated good performance across training languages and good generalization to unseen, related languages. This work examines whether, between related languages, multilingual models are equally *right for the right reasons*, i.e., if interpretability methods reveal that the models put emphasis on the same words as humans. To this end, we provide a new trilingual, parallel corpus of rationale annotations for English, Danish and Italian sentiment analysis models and use it to benchmark models and interpretability methods. We propose rank-biased overlap as a better metric for comparing input token attributions to human rationale annotations. Our results show: (i) models generally perform well on the languages they are trained on, and align best with human rationales in these languages; (ii) performance is higher on English, even when not a source language, but this performance is not accompanied by higher alignment with human rationales, which suggests that language models favor English, but do not facilitate successful transfer of rationales.

7.1 INTRODUCTION

NLP models are sometimes right for the wrong reasons, e.g., when sentiment analysis models correctly predict a movie review to be positive because it contains the word *Shrek* [204]. Human rationale annotations can be used to evaluate the extent to which models are right for the right reasons, i.e., whether model rationales align with human rationales. Datasets with rationale annotations exist for sentiment analysis [241], fact-checking [221], natural language inference

[1] In the published paper underlying this chapter, we did not consistently use the terminology regarding explainability as presented in Section 2.3, but loosely followed the style of Ribeiro, Singh, and Guestrin [187] and Lundberg and Lee [133].

EN	**A**	**deep**	**and**	**meaningful**	**film**
	2.34	1.69	2.70	1.92	0.09
DA	**En**	**dyb**	**og**	**meningsfuld**	**film**
	0.20	0.79	0.67	2.32	0.11
IT	**Un**	**film**	**profondo**	**e**	**significativo**
	0.44	0.28	1.72	1.79	1.43

Table 22: Tokens with machine generated importance scores for direct translations of the same sentence into English, Danish, and Italian. We see machine rationales are nevertheless quite different; e.g., consider the importance scores for the connectives *and*, *og* and *e*.

[32], and hate speech detection [140],[2] but so far only for the English language. While multilingual language models often fail to generalize across distant languages [169, 193, 205], they do bridge between related languages and have become a standard solution to data sparsity [246], as well as a way to reduce the overall energy consumption of training language-specific language models [194]. Benchmark performance does not tell us whether multilingual models are more prone to spurious correlations in some languages rather than others, i.e., whether models are *equally right for the right reasons* or to different degrees, see Table 22.

This paper presents a trilingual parallel corpus of human rationale annotations in Danish, Italian, and English, for the task of sentiment analysis. To this end, we translate an existing sentiment analysis dataset into different languages following a similar procedure as Hu et al. [91], with human post-correction. We then collect rationales from native speakers of these languages. We evaluate the quality of our human rationale annotations in two ways: using inter-annotator agreement metrics and using human forward prediction experiments [152]. We then use the corpus to evaluate the extent to which multilingual language models are equally right for the right reasons across languages, and whether agreement with human rationales aligns with downstream performance.

CONTRIBUTIONS Our contributions are as follows: (a) We present a trilingual corpus of human rationales, based on post-corrected translations of the Stanford Sentiment Treebank [209] and annotated by native speakers. The corpus is made publicly available at `https://github.com/RasmusKaer/BlackBox2022`. (b) We propose better metrics for comparing ranked rationales than has previously been used as well as a sequence-wise normalization of LIME's token scores to make scores comparable across sequences. (c) We evaluate mBERT [59] and XLM-R [46], in conjunction with two interpretability methods, LIME [187] and SHAP [133], across three languages, quantifying the extent to which these models are *equally right for the right reasons*.

[2] Several of these datasets can also be found in the ERASER Benchmark [55].

7.2 MULTILINGUAL RATIONALE ANNOTATION

Our multilingual corpus of human rationales is based on post-corrected translations of the Stanford Sentiment Treebank. We obtain Danish and Italian translations of a sample of validation data, correct the translations manually, and have native speakers annotate the original English sentences, as well as their post-corrected translations. We then validate the annotations by quantifying human inter-annotator agreement and by performing human forward prediction experiments [62, 79, 80, 86, 152]. We describe each step in detail in this section.

STANFORD SENTIMENT TREEBANK (SST) Our dataset builds on a sample of the Stanford Sentiment Treebank, which originally consists of 11,855 sentences from movie reviews, annotated with sentiment labels, and split in training, validation and evaluation sections of 8,544, 1,101, and 2,210 sentences. The sample selected for annotation of the rationales consists of 250 sentences from the validation section.

TRANSLATION We translate the English dataset into the target languages using Google Cloud API[3]. We carefully correct the translations of the rationales set manually and assess the quality of corpus through a language analysis. The post-correction process is presented in 7.6.3. We are aware that it would have been beneficial to have a set of languages that was more representative of linguistic diversity, but for this work we only had access to professional annotators in the three languages.

ANNOTATION We ask native speakers of English, Danish and Italian to annotate the sample with rationales. Our aim is to identify two types of information for each sentence: the rationales span, snippets of text that support the outcome; and the rank, the most meaningful words to justify the sentiment of the sentence. Inspired by previous explainability work in NLP using human rationale annotations [55, 140, 245], we follow the annotation guidelines in Zaidan, Eisner, and Piatko [242]. For the rank, we are interested in single words that carry a semantic meaning for the output (positive or negative sentiment). Annotators are asked to rank up to five words from most (1) to least (5) meaningful. See Table 23 for an example. The four annotators used in this study had linguistic training and participated on a voluntary basis.

S	John and Adam are such likeable actors.
R	John and Adam are such [2] likeable [1] actors.
S	A warm , funny , engaging film.
R	A warm [3], funny [1], engaging [2] film.

Table 23: Text annotation showing span (S) annotation and rank (R) annotation.

[3] Advanced version (v3), September 2021

ANNOTATOR AGREEMENT The inter-annotator agreement is measured as Cohen's κ [45] and accuracy; see Table 24. The κ coefficients suggest that the two annotators for each language have substantial agreement across all languages.

Lang.	κ	Acc.	Span	Rank	Tokens
DA	0.705	0.882	1,114	722	4157
EN	0.731	0.890	1,250	770	4232
IT	0.642	0.857	1,067	736	4411

Table 24: Annotator agreement and rationales by token. The minimum sentence length is 3 tokens for all three languages. The average length for both EN and DA is 17 and the maximum is 42 tokens per sentence, while in IT it is, respectively, 18 and 44 tokens per sentence.

FORWARD PREDICTION Besides calculating the inter-annotator agreement, we also validate the quality of our annotations through human forward prediction [62, 79, 80, 86, 152]. We recruited 9 annotators from our professional network, and everyone had degrees in computer science or linguistics. In a small-scale side experiment, we show participants 28 examples in which rationales identified by the annotators are highlighted. Participants are then asked to guess the ground truth (positive or negative sentiment) from these highlighted spans. We compare this to a baseline setting in which our participants have to guess the ground truth from raw text. We explicitly mentioned in the task that the results will be used for scientific research. If the rationales help participants predict the ground truth, they have been shown to be good rationales. Humans predicted the ground-truth for 82% of the examples with rationales, compared to 70% of the examples *without* rationales. For example, without rationales provided, 22.2% of annotators struggled in identifying the correct sentiment of a review such as *"Turns a potentially forgettable formula into something strangely diverting"*, while having less difficulties with equally challenging reviews when the rationales are provided. The high inter-annotator agreement and the usefulness of our rationales together indicate that our annotations are of high quality.

7.3 COMPARING RANKED RATIONALE LISTS

To evaluate the agreement between human rationales and rationales identified by interpretability methods applied to automatic sentiment analyses, we need a similarity measure for comparing ranked rationale lists. Common correlation tests are not sufficient, because the measure must be applicable to non-conjoint, uneven lists and should put a higher weight on higher-ranked words.

The human annotator selects the most relevant words in a sentence until exhausted. The ranking is ordered, but may only contain a few words. On the other hand, the interpretability methods provide by design a rank for each word in a sentence. Thus, the annotator's ranking is typically *incomplete* (not all items are ranked), while the automatically computed ranking is *complete*. That is, the two rankings are mutually *non-conjoint*. Furthermore, we need to deal with *indefiniteness* [229] in the sense that the annotator may truncate the complete list at an arbitrary depth. The measure we propose for evaluating rationale rankings is the extrapolated version of the *rank-biased overlap* [229], RBO_{EXT}, which is a generalization of average based overlap for indefinite rankings. It ranges from 0 (disjoint) to 1 (identical). The RBO_{EXT} measure satisfy the criteria needed for evaluating the agreement of list rationale rankings of both sentences and documents by being able to handle tied ranks, rankings of different lengths and top-weighted rankings.

The degree of top-weightedness is determined by a parameter $p \in [0, 1]$. Consider a person comparing two rankings by sequentially going through the lists starting with the highest rank. In each step, one additional rank is considered. That is, in the beginning only the highest ranked elements are compared, then additionally the top two elements are compared, and so on. At each step, the person stops the comparison with a probability $1 - p$. Roughly speaking, RBO_{EXT} measures the expected similarity computed by this randomized comparison. The parameter p induces a weighting of the ranks that decreases with decreasing rank (i.e., decreasing importance). Following Webber, Moffat, and Zobel [229], we choose p such that 86% of the weight is concentrated on the first d ranks. They show that the concentration of weights on the first d ranks given p can be computed as

$$1 - p^{d-1} + \frac{1-p}{p} d \left(\ln \frac{1}{1-p} - \sum_{i=1}^{d-1} \frac{p^i}{i} \right).$$

Table 24 shows that annotators on average rank 3 words per sentence. Hence, we set $p = 0.68$, because this leads to a concentration of roughly 86% for $d = 3$. The annotators were asked to rank up to 5 words. Therefore, we also considered only the top-5 elements in the rankings produced by the interpretability methods (still, we apply RBO_{EXT} as derived for indefinite rankings).

7.4 EXPERIMENTS

Our experiments below rely on two pretrained multilingual language models, which we briefly introduce, three different experimental protocols, and two different interpretability methods.

7.4.1 *Pretrained language models*

The experimental protocol is based on two pretrained multilingual transformer language models [224], namely mBERT [59][4] and XLM-R [46][5]. We used the base, cased version from the Hugging Face trans-

[4] https://huggingface.co/bert-base-multilingual-cased
[5] https://huggingface.co/xlm-roberta-base

formers library[6]. Following Devlin et al. [59], we added a classification layer on top of the [CLS] token. We fine-tuned these models for 3 epochs on a single Tesla K80 GPU, with a training batch size of 16 and a learning rate of $3 \cdot 10^{-5}$. The parameters were found using manual hyperparameter tuning based on the authors' recommendations of batch-sizes $\{16, 32\}$, epochs $\{2, 3, 4\}$. The learning rate was fine-tuned over $\{2 \cdot 10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$ with 3 trials each.

7.4.2 *Experimental protocols*

In our experiments, we fine-tune mBERT and XLM-R on the SST training data and/or translations thereof (into Danish or Italian). We rely on three standard protocols, which we call the BASE-SETTING, the CROSS-SETTING, and the MULTI-SETTING. In the BASE-SETTING, we fine-tune mBERT and XLM-R on a single language, e.g., English, and evaluate them on the evaluation data in the *same* language. This corresponds to the situation in which you use a multilingual language model to learn a monolingual model in the presence of training data. This scenario is common for *medium-resourced* languages. In the CROSS-SETTING, we evaluate such models, e.g., trained on English, on another language. This scenario is common for *low-resourced* languages. Finally, in the MULTI-SETTING, we train and evaluate on all three languages, inducing a *multilingual* sentiment analysis model for three languages. In all three settings, we evaluate the extent to which the fine-tuned mBERT and XLM-R models align with human rationales, relying on interpretability methods.

7.4.3 *Interpretability methods*

A variety of methods for deriving explanations are currently being used by the NLP community. Examples of such methods are LIME [187] and SHAP [133], LRP [14], and DTD [147]. For this study, we consider SHAP and LIME, since they are two of the most widely used post-hoc model interpretability methods, also used in similar studies such as ERASER [55] and HateXplain [140]. LIME is a model-agnostic approach that returns an explanation for a prediction on an input example (a text) by virtue of a local linear approximation of the model's behavior around that example. The linear approximation is a sparse linear model induced from hundreds of perturbations of the example. In the case of text examples, perturbations are obtained by randomly removing tokens or words. SHAP is also model-agnostic and based on Shapley values [201], a concept from cooperative game theory, which refers to the average of the marginal contributions to all possible coalitions. When applied to text, the method, like LIME, produces explanations in terms of tokens or words. We kept the hyperparameters of the two methods to their default-setting, except for the size of neighbourhood used to learn linear models for LIME, which we set to 500 for computational reasons.

[6] https://huggingface.co/docs/transformers, V4.15.0

7.5 RESULTS

Protocol settings				SHAP		LIME	
Source	Model	Target	Acc.	ROC AUC	RBO_{EXT}	ROC AUC	RBO_{EXT}
English	EN-mBERT	EN	81.48 ± 0.3	68.69 ± 0.7	51.63 ± 0.0	67.08 ± 0.0	53.76 ± 0.0
		IT	74.28 ± 0.6	70.11 ± 1.0	49.92 ± 0.0	66.18 ± 0.0	47.77 ± 0.0
		DA	70.42 ± 0.9	67.41 ± 1.0	44.38 ± 0.0	62.05 ± 0.0	42.35 ± 0.0
	EN-XLM-R	EN	85.37 ± 0.2	69.95 ± 1.4	52.78 ± 0.0	66.83 ± 0.0	56.87 ± 0.0
		IT	82.16 ± 0.2	69.80 ± 0.4	48.52 ± 0.0	68.05 ± 0.0	54.48 ± 0.0
		DA	82.50 ± 0.3	68.85 ± 0.7	50.68 ± 0.0	66.19 ± 0.0	53.33 ± 0.0
Italian	IT-mBERT	IT	80.66 ± 1.2	69.24 ± 1.1	53.24 ± 0.0	68.23 ± 0.0	55.37 ± 0.0
		EN	76.08 ± 1.7	68.79 ± 1.0	50.46 ± 0.0	66.04 ± 0.0	48.62 ± 0.0
		DA	68.94 ± 0.5	65.13 ± 0.6	43.11 ± 0.0	62.66 ± 0.0	43.95 ± 0.0
	IT-XLM-R	IT	82.56 ± 0.0	71.79 ± 1.2	52.79 ± 0.0	69.94 ± 0.0	56.72 ± 0.0
		EN	84.15 ± 0.7	70.62 ± 0.8	55.48 ± 0.0	66.79 ± 0.0	55.22 ± 0.0
		DA	81.24 ± 1.0	69.59 ± 0.4	53.03 ± 0.0	66.16 ± 0.0	52.98 ± 0.0
Danish	DA-mBERT	DA	79.17 ± 0.5	67.40 ± 2.0	49.07 ± 0.0	66.37 ± 0.0	51.33 ± 0.0
		IT	72.10 ± 0.3	68.36 ± 0.8	45.84 ± 0.0	64.74 ± 0.0	45.39 ± 0.0
		EN	75.60 ± 0.7	69.95 ± 0.5	49.50 ± 0.0	66.17 ± 0.0	48.37 ± 0.0
	DA-XLM-R	DA	83.41 ± 0.5	69.74 ± 1.6	55.88 ± 0.0	65.99 ± 0.0	53.27 ± 0.0
		IT	82.07 ± 0.6	69.16 ± 0.6	49.75 ± 0.0	67.57 ± 0.0	52.12 ± 0.0
		EN	84.80 ± 0.2	70.39 ± 1.1	53.63 ± 0.0	66.34 ± 0.0	52.59 ± 0.0
Multi	MULTI-mBERT	EN	81.51 ± 0.1	65.02 ± 2.1	43.49 ± 0.0	65.97 ± 0.0	51.68 ± 0.0
		IT	80.62 ± 0.2	66.16 ± 1.6	45.57 ± 0.0	66.21 ± 0.0	49.60 ± 0.0
		DA	78.34 ± 0.9	63.99 ± 0.4	42.65 ± 0.0	63.89 ± 0.0	49.71 ± 0.0
	MULTI-XLM-R	EN	85.83 ± 0.4	67.79 ± 0.8	50.45 ± 0.0	64.48 ± 0.0	48.66 ± 0.0
		IT	83.67 ± 0.3	69.10 ± 0.7	46.41 ± 0.0	66.52 ± 0.0	51.88 ± 0.0
		DA	82.88 ± 0.7	66.99 ± 1.3	48.89 ± 0.0	64.61 ± 0.0	49.59 ± 0.0

Table 25: Evaluation results on the multilingual corpus of rationales. All results are averaged over three trials. We report the results in percentages. We observe that generally models perform well on the languages they are trained on (source languages), and align best with human rationales in these languages. Generally, mBERT aligns better with human rationales, but XLM-R performs better. We also observe, however, that performance is high on English, even when not a source language, but that this performance is not accompanied by higher alignment with human rationales. This suggests that language models favor English, but do not facilitate successful transfer of rationales.

Table 25 presents the results of the experimental protocol on our trilingual corpus. We compare the effectiveness of LIME and SHAP on human rationales. The agreements is evaluated using ROC AUC for rationale span and RBO_{EXT} for rank similarity based on all 250 samples. The protocol sets two properties for fine-tuning: a single language, denoted by DA, EN and IT, or multiple languages, denoted

MULTI. The fine-tuned models are tested across DA, EN and IT with 3 runs per setting.

PERFORMANCE OF MBERT AND XLM-R The accuracy of the multilingual models across languages and settings is presented in Table 25. The results confirm the findings of the original works [46], that XLM-R is consistently better than mBERT.

While mBERT-based models consistently obtain their highest accuracy in the BASE-SETTING, XLM-R-based models always perform best on English as the target language, independently from the source language. mBERT-based models exhibit a high variation in the CROSS-SETTING (5.11 p.p. difference between the average accuracy of the BASE compared to the CROSS settings), e.g., EN-mBERT achieves 81.48% accuracy when tested on the English test set, but has only 70.42% accuracy on Danish. In contrast, XLM-R shows less variation between BASE and CROSS settings (0.52 p.p. difference).

But does a higher performance correspond to higher agreement with human rationales? Table 25 presents the results for agreement, evaluated using ROC AUC for rationale span and RBO_{EXT} for rank similarity of the two list rankings. The results suggest that the accuracy of the models does not generally seem to influence ROC AUC and RBO_{EXT} scores, since a much higher accuracy does not imply better span prediction.

INTERPRETABILITY METHODS Our evaluation of the span agreement shows an average across all models and languages of 68.50% for SHAP and 66.04% for LIME, indicating that SHAP has a higher (2.46 p.p.) agreement with human span rationales than LIME. The average rank agreement across all models and languages measured using RBO_{EXT} is 49.46% for SHAP and 51.07% for LIME, the latter being 1.61 p.p. higher in agreement than SHAP. These experiments show that we do not have a single best method across rank and span. Our results suggest a trend of SHAP being a more successful method for capturing good weights for span agreement and LIME being slightly more in accordance with human ranking.

LANGUAGES The best rank agreement is achieved when English is used as target language, with the overall highest for both LIME (51.97%) and SHAP (50.93%), as presented in Table 26.

The second best rank agreement is obtained in Italian, while the worst is in Danish for both LIME and SHAP. The highest average span score is achieved on Italian, while English follows close and Danish again remain the lowest in agreement. While English is slightly higher in rank agreement, Italian obtains a better span agreement. The lowest span and rank agreement is generally seen with Danish as target language. As we are interested in how languages compare across models, settings and metrics, we can derive the total from the target languages column in Table 26. Altogether, these results indicate that we have better explanations for English (59.50%) than we have for Italian (59.27%) and Danish (57.54%). The explanations for

Metric	Method	Target-EN	Target-IT	Target-DA
RBO_{EXT}	SHAP	50.93	49.01	48.46
	LIME	51.97	51.67	49.56
ROC AUC	SHAP	68.90	69.22	67.39
	LIME	66.21	67.18	64.74
Overall		59.50	59.27	57.54

Table 26: To investigate whether explanations are in equal agreement across languages, we group target languages together across the BASE, CROSS and MULTI settings.

English are 1.96 p.p. higher in agreement with human rationales than the explanations derived from Danish, while Italian is 1.73 p.p. higher than Danish.

EVALUATION METRICS An interpretation of the evaluation metrics across settings and languages shows a span agreement that ranges from 62.05% to 71.79%, with an average of 67.27%. What we can interpret from the score is a satisfactory span agreement, suggesting that there is a $\frac{2}{3}$ chance that the model is able to distinguish a token inside a span and a token outside a span. That is, the machine rationale agrees with a human rationale. Regarding the rank agreement across all settings and languages, we see it ranges from 42.35% to 56.87% with an overall average of 50.27%. The score can be interpreted as neither disjoint nor identical, thus implying a fair agreement.

7.6 ANALYSIS

In this section, we present our analysis of our results and findings. First, we address whether models are *equally right for the right reasons* and how performance compares to agreement. Next, we analyze the translations and the post-corrections. Lastly, we examine whether token scores predict human rationales.

7.6.1 *Are models equally right for the right reasons across languages?*

The idea of being right for the right reasons refers to learning from reliable signals in your data, which are causally related to the ground truth classification. While some models can be used to illuminate complex causal dynamics, others adapt Clever Hans strategies of relying on pervasive, yet spurious correlations in the training data. In this paper, we ask if multilingual language models such as MBERT and XLM-R are equally prone to spurious correlations across languages? Or could it be that these models adopt Clever Hans strategies for some languages, but not for others?

Our results show, very consistently, that MBERT and XLM-R are *less* right for the right reasons for Danish: When the training language is English or Italian, or when multilingual training language is used, Danish never aligns best with human rationales. For English and Italian, it comes in worst in 18/20 cases, and in the multilingual setting, Danish is least right for the right reasons in 6/10 cases. For English and Italian, things are more or less *on par*. While English is slightly higher in rank agreement, then Italian obtains a better span agreement, but the lowest span and rank agreement is generally seen with Danish as the target language. We conclude that multilingual language models are *not* equally right for the right reasons across languages.

7.6.2 *How indicative is accuracy for agreement?*

It seems intuitive that a good model with high performance will also align better with human rationales, but theoretically, models may adopt radically different strategies, if multiple strategies are possible. Even if we expect a positive correlation between performance and alignment, how strong is this correlation in practice? To answer this question, we compute the correlation between the accuracy of the language models and the agreement of span and rank. We use Spearman's rank-order correlation test and Pearson's correlation test, across both explanation methods and all datasets. Both tests show that performance is only weakly (positively) correlated with alignment with human rationales; see Table 27 for details. That is, we see better alignment if models are better, but performance explains only a little of the variance, suggesting multiple possible strategies for prediction exist. This aligns well with our results, also, where a larger difference in accuracy between models does not transfer into a significant difference in agreement.

Lang.	Spearman's ρ	Pearson's ρ
Acc/AUC	0.059^{**}	0.092^{**}
Acc/RBO	0.076^{**}	0.153^{**}

Table 27: Correlation scores for performance (Acc) and alignment with human rationales (AUC/RBO).

Humans may base their rationales on different parts than machine-based rationales. While humans consider *and* necessary for the snippet of *deep and meaningful* (see example in Table 22), a model may not find it a useful predictor of sentiment. Humans and models may agree on the sentiment, but for slightly different reasons.

7.6.3 *Language analysis*

The translated corpus is post-corrected to obtain a high overall quality, ensuring that the corpus can be used to evaluate the interpretability methods in our experiments. To quantify the translations quality, we report the number or sentences that needed corrections and the average number of corrected words in Table 28. The percentage of

Lang.	% corrected sentences	Avg. corrected words
DA	15.60	1.46
IT	17.20	1.74

Table 28: Percentage of corrected sentences and average number of corrected words per sentence in Italian and Danish.

sentences that needed to have corrections in Italian and Danish are, respectively, 17.20% and 15.60%. Among these corrected sentences, 1.74 words were corrected on average in Italian, 1.46 in Danish. The results indicate that overall the quality of the translations is high. This is also supported by the performance of the fine-tuned models in Table 25. A selection of original translation and the post-corrected equivalent is presented in Table 29. We can highlight some limitations found during post-correction. The original sentences sometimes present an informal register, sprinkled with colloquial and slang words, which may result in suboptimal and literal translations. Some of the original sentences present idiomatic expressions that might result in a literal translation, as in A-DA, not corresponding to actual terms in the target language. Moreover, some translations may contain subpar syntactic structure or lexicon, e.g., in A-IT *brutto* is more suiting to refer to *films*, although it presents the same polarity and magnitude of the original adjective. In B-IT the sentiment of the expression could be misinterpreted, since *fa impazzire* is sometimes used in a positive connotation. Lastly, sometimes the original English sentences contain typos and other errors, which the model is understandably not able to correct or process, therefore transferred into the translations.

A-IT ORG.	..., sbalorditivo, assurdamente *cattivo.*
A-IT COR.	..., sbalorditivo, assurdamente **brutto**
B-IT ORG.	Questo film *fa impazzire.*
B-IT COR.	Questo film **è esasperante**.
A-DA ORG.	Der er *parcelhuller*, der er store nok til, ...
A-DA COR.	Der er **plothuller**, der er store nok til, ...
B-DA ORG.	Det er en *greb taske* med genrer, ...
B-DA COR.	Det er en **rodekasse** med genrer, ...

Table 29: Examples of corrected translations (COR.) and the original translations (ORG.).

7.6.4 *Do token scores predict human rationales*

Meaningful token scores produced by an interpretability method should be predictive of human rationales [55, 62, 152]. To verify this, we map the token score $s(w)$ of a word w to an estimate of the probability that the word is in the rationales span. We assume a logistic model

$$P(w \text{ in rationales span} \mid s(w)) = \sigma_{a,b}(|s(w)|) \enspace ,$$

where $\sigma_{a,b}(x) = (1 + \exp(ax + b))^{-1}$ with scalar parameters a and b. These parameters are determined by maximum likelihood estimation on a training set pairing token scores and corresponding human annotations. We consider the absolute value of the score because we are interested in the importance of a word regardless of whether it contributes to a positive or negative sentiment. This approach corresponds to calibrating the (absolute) scores to posterior probabilities as suggested by Platt [153, 171]. It can also be viewed as logistic regression from the absolute score to the dependent variable indicating whether a word is in the rationale span or not.

The logistic model gives us the probability of a word being a rationale, which allows for an interpretation of token scores and a comparison of scores across different interpretability methods. In particular, the model suggests a criterion for deciding whether a word should be considered part of the rationales span or not by applying the natural 50% threshold on the probabilities (we pay for this additional information by using training data to fit the models). To fit the model and to compare the different interpretability methods, we split our data into a training and a validation set. We used 25 positive and 25 negative samples for validation and trained on the remaining 200 data points.

Let $\mathbf{s} = (s(w_1), s(w_2), \dots)^T$ denote the vector of scores for a word sequence $w_1, w_2, \dots$ and $\min(\mathbf{s})$ and $\max(\mathbf{s})$ the minimum and maximum element of $\mathbf{s}$, respectively. To compare token scores across sequences, their scaling should not differ across the sequences. That is, because we can assume that each sequence contains at least one word within and one outside the span, for two sequence $\mathbf{s}$ and $\mathbf{s}'$ we should have $\min(\mathbf{s}) = \min(\mathbf{s}')$ and $\max(\mathbf{s}) = \max(\mathbf{s}')$. We found this property to be violated, in particular for LIME. Thus, we normalized the scores at the sequence level using

$$s(w) \leftarrow \frac{s(w) - \min(\mathbf{s})}{\max(\mathbf{s}) - \min(\mathbf{s})}$$

for each score $s(w)$ in a sequence with scores $\mathbf{s}$.

Table 30 shows the accuracies on the held-out sets in BASE-SETTING. Both methods performed better than simply predicting the majority class. Without normalization, SHAP outperformed LIME on our (rather small) validation data set. LIME was only slightly better than the baseline, but after normalization LIME surpassed SHAP, which did not profit from the normalization. When evaluating explanations on how well the token scores generalize to human rationales, we see a similar pattern of Italian and English sharing the highest agreement where Danish consistently shows the lowest agreement.

Human annotated rationales include connectives, determiners, and similar, which are irrelevant for our binary task and are therefore not used by the logistic models. This suggests that methods for adding

		LIME	LIME	SHAP	SHAP	BASE
		mBERT	XLM-R	mBERT	XLM-R	LINE
(A)	EN	70.03	71.51	71.68	72.76	67.74
	DA	69.50	70.23	70.83	72.75	67.49
	IT	70.94	72.73	72.73	73.78	67.80
(B)	EN	73.75	73.03	70.97	71.68	67.74
	DA	72.34	72.75	71.47	72.70	67.49
	IT	73.47	75.44	73.30	73.13	67.80

Table 30: The accuracies on the hold-out sets in BASE-SETTING. The BASELINE is a majority classifier that naively predicts all tokens as not a rationale. (A) refers to the original token scores and (B) to the normalized token scores.

the relevance of these could be a promising direction for improving our approach and the evaluation between human and machine rationales.

7.7 RELATED WORK

Transformer-based multilingual models have been analyzed in many ways: Researchers have, for example, looked at performance differences across languages [205], looked at their organization of language types [182], used similarity analysis to probe their representations [113], and investigated how learned self-attention in the Transformer blocks affects different languages [184]. Human rationales have been used to supervise attention for various text classification tasks, such as sentiment analysis [247] and machine translation [240]. Feature attribution methods such as LIME and SHAP have also been applied to multilingual models: LIME has been applied to mBERT for analysis of hate speech models [7], and SHAP has been applied to mBERT in biomedical NLP [243]. LIME has also been applied to XLM-R in the context of hate speech [208], as well as in a biomedical context [110]. Shapley values have also been used to estimate the influence of source languages on the final predictions of models based on mBERT [160]. None of these applications have been evaluated, however. Feature attributions have been applied to monolingual models, especially for English, more often than multilingual models. For English, we have a set of datasets with human rationales that we can use to evaluate feature attribution methods. These include BeerAdvocate [15] and e-SNLI [32], as well as other datasets, several of which were collected in the ERASER benchmark [55]. The reason feature attribution methods have not been properly evaluated in a multilingual context, is simple: There was, until now, no gold standard with which to evaluate the rationales produced by multilingual models.

7.8 CONCLUSIONS

We introduced a new trilingual, parallel corpus of human rank and span rationales in three related languages, English, Danish and Italian. We proposed rank-biased overlap as a better metric for rank evaluation when common correlation tests are not sufficient. We found that a sequence-wise normalization of LIME's token scores is required to make scores comparable across sequences. Evaluations on the corpus showed that generally, models perform well on the languages they are trained on, and align best with human rationales in these languages. Models can be right for different reasons. The main results suggest that multilingual models are *not* equally right for the right reasons in the sense that interpretability methods indicate that the models not necessarily put emphasis on the same words as humans. We also observed that performance is high on English, even when it is not a source language, but that this superior performance is not accompanied by higher alignment with human rationales. In other words, this zero-shot advantage of English as a target language seems to come at the cost of being more prone to spurious correlations. With this work, we hope to inspire further progress on multilingual interpretation and collection of rationales in different languages.

LIMITATIONS

All the languages chosen for the presented work belong to the Indo-European language family, since we only had access to professional annotators in the three languages. A clear limitation of this study is the lack of linguistic diversity in the set of languages used. It would be beneficial in the future to build larger rationale datasets for less related languages, including languages from different language families. Another limitation to be highlighted is the limited size of the multilingual parallel corpus of rationales, consisting on 250 annotations per language. Finally, although the parallel corpus was post-corrected, the language models are fine-tuned on the translations.

Part V

CONCLUSION

8 DISCUSSION AND CONCLUSION

The preceding chapters presented new work in three research areas with an emphasis on financial NLP. In this concluding discussion, the studies presented will be reviewed in the context of the general focus of this dissertation on advancing natural language processing for applications in the financial domain.

Part II on Financial Transactions presented machine learning-based systems for the classification of financial transactions. In chapter 4, machine learning systems were devised to automate the accounting task. The results showed an average accuracy above 80% when considering 473 companies individually with a per company classifier. We found that processing transaction texts considerably improved the performance, where using word embeddings with subword information outperformed the baseline of a lexical bag-of-words representation. After the unification of account structures and feature engineering, the system generalizes across companies and different corporate sectors. We trained a single classifier on 44 companies belonging to 28 different sectors. It achieved high performance across companies and corporate sectors, even on new companies with no historical data.

We received feedback from eight accounting and bookkeeping experts, who were satisfied with the high accuracies and the advances that the proposed systems demonstrated. They estimated these solutions could improve quality and save approximately 30-50 percent in costs and time compared to current rule-based systems.

In terms of avenues for further work, one direction would be to ease the multiple charts of accounts problem by designing templates that enable all companies to convert from individual charts to a unified chart. Another direction would be to examine embeddings trained on domain-specific transaction text and multilingual embeddings.

Part III on Multilingual Financial NLP advances multilingual NLP in the financial domain. In chapter 5, we studied multilingual domain-specific models, focusing on adapting a single model to multiple languages within a specific domain, in particular the financial domain. To address the issue of most existing work being mainly English-based, we achieved multilingual domain adaptive pretraining (mDAPT) in a single model by extending domain adaptive pretraining to a multilingual scenario. The study proposed different techniques and strategies for making a single model become both domain-specific and multilingual through different settings of pretraining datasets for continued pretraining of language models. We considered the limitations of data and computational resources by placing a budget on the size of the resulting corpora. Further, we explored up-sampling using general-domain data and employed both adapter-based and complete model pretraining for mDAPT.

The results showed that the proposed multilingual domain-specific model could outperform the general-domain multilingual model and come close to its corresponding monolingual counterpart. Having a single model instead of a distinct model per language eases deployment and demands comparably less computational resources. The results hold across different domain-specific datasets representing seven languages and two pretraining methods. This work also confirmed the findings of Gururangan et al. [83] and Araci [8], and underlines the importance of domain adaptation to better address domain-specific tasks. The mDAPT models are publicly available to practitioners and the research community.

An interesting future study would be to assess which pretraining methods and models are better suited for being adapted to a specific domain. Not only with respect to the performance of the methods but also concerning the required resources. As emphasized previously in chapter 2, training longer over more data with a larger number of model parameters may further improve the performance of domain-specific models, given that a larger multilingual pretraining corpus can be collected. As increasing the model size tends to improve performance, it also sets a greater demand on both data and computational resources, posing a potential limitation for many institutions and practitioners. On the other hand, carrying out future work on more effective methods that use smaller models, including model distillation, could also be a promising direction for further work and a step towards more resource-efficient research.

Chapter 6 proposed a benchmark dataset for the evaluation of multilingual financial language models. The financial benchmark contains 10,000 examples, covering 15 languages, including different language families and writing systems. We benchmarked popular generic multilingual language models on the Multilingual Financial Benchmark and found that these models generally perform well in high-resource languages but present a performance gap between high- and low-resource languages in the benchmark. We further analyzed the benchmarked models through a per-language analysis, which identified that most models struggle to facilitate a good transfer across the evaluated languages. This also revealed a substantial need for improving the models' capacity to generalize to specific languages. We compared the general-domain models against a domain-adapted counterpart for which we used the mDAPT model presented in chapter 5. The multilingual domain-adapted model demonstrated better generalization across the evaluated languages and performed much better on low-resource languages than the general-domain models. It suggests that domain-specific models and methods are promising directions in financial NLP. In addition, we also presented a multilingual pretraining corpus of financial texts in 14 languages (FinMultiCorpus) and a non-English sentiment dataset (DanFinNews) to support work on multilingual NLP in the financial domain.

Regarding the financial benchmark, future work should focus on extending the dataset with more examples across more languages and investigating an extension with complete articles for document processing. These initiatives would help community to understand and explore the limits of multilingual NLP in the financial domain.

Part IV on EXPLAINABILITY IN MULTILINGUAL NLP evaluates the explanations produced by explainability methods for multilingual NLP systems. We found motivation for explainable NLP systems when used across languages from our work on multilingual NLP in the financial domain. Although this interest evolved in relation to the financial domain, we pursued this work on general text so that we could build on current explainability research and make the data publicly available.

Chapter 7 analyzed whether comparable performance figures can be observed or if severe robustness gaps are hidden between related languages. The results showed, on the provided parallel corpus, that multilingual models perform better on languages seen during fine-tuning, although the unseen languages are part of the pretrained languages. The alignment with human rationales was also better for those languages. However, it was also observed that performance on the English language is high even when not seen during fine-tuning. This suggests that language models favor English and that high accuracy does not necessarily lead to a more successful transfer or a higher alignment with human rationales. The investigation presented also suggested rank-biased overlap as a more suitable metric for rank evaluations and a sequence-wise normalization of LIME's token scores. This study provided a trilingual parallel corpus of human rationale annotations in Danish, English, and Italian to benchmark models and explainability methods.

Chapter 7 confirmed the findings by Atanasova et al. [11] who found SHAP outperformed LIME with respect to span agreement with human rationales. We extended this observation to multiple languages. As mentioned in chapter 7, span rationales contain words such as connectives, determiners and modifiers of low semantic saliency that may not be necessary for the model to correctly predict a binary sentiment. Thus, a model for predicting binary sentiment can perform well if these words are not assigned a high enough weight by the explainability methods to be considered [100] – one could even argue that a good model should not be influenced by these words. Thus, comparing the input words important for a model one-to-one with human span rationals in the context of binary sentiment prediction may not be a very good indicator for model performance. Therefore, we extended the study also to include rank agreement in the evaluation. The rank agreement takes the importance of the individual words into account, putting less emphasis on words with low semantic saliency. Surprisingly, the results showed that LIME performed slightly better than SHAP with respect to rank agreement with human ranking. That is, our approach for comparing agreement, which addresses issues with using one-to-one comparisons with human span rationals, could suggest a reevaluation of explainability methods. Further research should be undertaken to investigate this difference, the methods and the settings using rank and span.

The study in chapter 7 was limited to a single dataset. In the future, we would like to consider other datasets, preferably multilingual datasets. From my perspective, the next step would be to expand the work into the financial domain. One direction would be to expand the MultiFin dataset with documents and annotate rationales for a subset of these documents. Another interesting study would be to examine the sentiment of financial text with a focus on negative words, as the tone in financial text can be different from the tone in the general domain [128]. General directions for future work could find inspiration in the multilingual benchmarks, XTREME [91] and XTREME-R [192], and consider collection of rationales in different languages and tasks.

With this work, the hope is to inspire further progress in explainability in multilingual NLP, as the explainability of multilingual NLP systems is an important area that calls for more work. As pointed out in part I, this is particularly important as human end-users may rely on the explanations produced by these explainability methods [55, 133, 187, 210], also in multilingual environments.

Part VI

APPENDIX

A SUPPLEMENTARY MATERIAL FOR INDIVIDUAL STUDIES

A.1 CHAPTER 4

A.1.1 *Using PCA to reduce the dimensions of the learned embeddings*

Because of limited training data and resources, we rely on pretrained word embeddings. The pretrained word vectors were learned on general data (Wikipedia and Common Crawl), have a dimensionalty of 300, and are available for multiple languages [22, 81, 104].

In Scenario II, we encode the three text fields Transaction Text, Subject Sector Text and External Sector Text, which results in 900 input features when using the embedding directly. This dimensionality is rather high given the amount of our financial training data. In particular, the baseline nearest neighbor classifier is known to suffer from the "curse of dimensionality" in the sense that "in high-dimensional feature spaces, more training data may be required to see enough combinations of different feature values appearing" [40].

A dimensionality of 900 is also high compared to the number of other input features. There are only few other features such as the transferred amount. Thus, there is a risk that the non-text features are overshadowed by the text features. This has an obvious effect on the random forest training, where the probability of selecting a non-text feature for splitting at the nodes of the decision trees gets small if to many text features are considered.

For these reasons, we looked at further reducing the dimensionality of the text embedding. The 300 dimensions are the result of learning a latent representation for general text. We used principal component analysis (PCA) to further reduce the dimensionality. However, a PCA extracting the most important dimensions for general text may remove particular information specific for financial texts. Therefore, we conducted a PCA on the 300-dimensional embeddings of our financial training data to extract the principal components of transaction texts (the "finance subspace").

Table 31 presents the results of the prior investigation to Scenario I leading to the experiments in sec. 4.4.3, Table 7. The tables shows how often how many principal components were selected when optimizing the OOB error of random forests trained for individual companies. In most cases a rather low dimensionality was selected. The results suggest that considering 10 components is a reasonable choice when using a single fixed number of components for all companies. However, the selected number of components varied across companies. This motivated us to include the number of components in the model selection in some of our experiments.

P.C.	# Comp.	Pct. (indiv.)	Pct. (acc.)	# Comp.	Pct. (indiv.)	Pct. (acc.)
	FastText			*BoW*		
1	9	1.90%	1.90%	10	2.11%	2.11%
2	6	1.27%	3.17%	13	2.75%	4.86%
4	78	16.49%	19.66%	83	17.55%	22.41%
8	105	22.20%	41.86%	84	17.76%	40.17%
16	114	24.10%	65.96%	113	23.89%	64.06%
32	114	24.10%	90.06%	105	22.20%	86.26%
64	40	8.46%	98.52%	58	12.26%	98.52%
128	7	1.48%	100.00%	7	1.48%	100.00%

Table 31: Selected components from setting *complete feature set (A)* with varying components from Table 7 using Random Forest. P.C. refers to the number of principal components for the word embedding by choosing $d_{PC} \in \{2^i \,|\, i = 0, \ldots, 7\}$ for each company individually. For each number P.C. of components and embedding method, the tables shows the absolute and relative number of companies for which P.C. components were selected in the columns # Comp. and Pct. (indiv.), respectively, as well as the relative number of companies for which P.C. or less components were selected in column Pct. (acc.).

A.1.2 *Baseline classifiers*

Across the experiments presented in Tables 7, 8 and 9, we observe that overall *k*-Nearest Neighbor performed better in 8 of the 10 experiments compared to the Logistic Regression. This indicates that the transaction data requires non-linear machine learning. The results of the baseline methods are given to assess the difficulty of the task and to benchmark the performance of our system. In addition, the base classifiers also confirm that random forest generally gives good results in practice [69].

A.1.3 *Choice of evaluation metric*

The choice of using classification accuracy as the metric to evaluate the classification of transactions is because false positives and false negatives have the same costs. In this domain, a wrong prediction has the same cost regardless of the class, as it would require a manual correction.

A.1.4 *Transformer models and FastText*

The study was not about examining the best way to represent the transaction text but showing that making use of the transaction text indeed helps in the classification, and this hypothesis is clearly supported [102].

The choice of a simpler vector embedding at the time of conducting the study in contrast to more complex transformer models is justified by the task and the available resources. An embedding for Danish was needed. The short text fields of the financial transactions do typically not contain complete sentences, thus there seems to be no need for large transformer models that produce a contextual embedding. The high efficiency of the solution, e.g., fast execution times, not requiring GPUs, and accordingly, small carbon footprint is an advantage.

A.1.5 *Using geolocational information for feature engineering*

The motivation for using geolocation information originated in the observation that invoices typically use high-level company names, but these can have different addresses depending on the subdivision of the company. For example, a company in telecommunication can have various subdivisions for handling cellphones, Internet, private customers, industrial customers, etc. Simply using the high-level company name does not reveal the subdivisions. Encoding the distance allows to distinguish between different subsidiaries. In addition, the distance may be related to different types of interactions, for example, may indicate local and more global interactions.

Table 10 presented the feature importance from experiment A.I, showing that the Distance feature is the second most important feature not originating from the text. This confirms the hypothesis that distinct distances for the subsidiaries result in a good predictive feature.

A.2 CHAPTER 5

A.2.1 *Baseline models*

Table 32 is a comparison between baseline `mono` models and the `multi` model. For the NER tasks, we use the `cased` versions for all experiments. For sentence classification, we use `uncased` versions for DA-BERT and EN-BERT.

A.2.2 *Biomedical data*

PREPROCESSING PRETRAINING DATA For the English abstracts, we sentence tokenize using NLTK and filter out sentences that do not contain letters. For the WMT abstracts, we filter out lines that start with #, as these indicate paper ID and author list. We determine the language of a document using its metadata provided by PubMed. We transliterate Russian PubMed titles (in Latin) back to Cyrillic using the `transliterate` python package (`https://pypi.org/project/transliterate/`).

DOWNSTREAM NER DATA The French QUAERO [155] dataset comprises titles of research articles indexed in the biomedical MEDLINE database, and information on marketed drugs from the European Medicines Agency. The Romanian BIORO [145] dataset consists of

	Training data	Vocab size	# parameters
de-BERT [36]	OSCAR (Common Crawl), OPUS (Translated web texts), Wikipedia, Court decisions [163.4G]	30.0K	109.1M
da-BERT	Common Crawl, Wikipedia, Debate forums, OpenSubtitles [9.5G, 1.6B]	31.7K	110.6M
en-BERT [59]	English Wikipedia, Books [3.3B]	29.0K	108.3M
en-bio-bert [120]	Initialized with en-BERT; continue on PubMed, PMC [18B]	29.0K	108.3M
FinBERT [8]	Initialized with en-BERT; continue on News articles [29M]	30.5K	109.5M
es-BERT [34]	OPUS, Wikipedia [3B]	31.0K	109.9M
fr-BERT [117]	24 corpora, including CommonCrawl, Wikipedia, OPUS, Books, News, and data from machine translation shared tasks, Wikimedia projects [71G, 12.7B]	68.7K	138.2M
pt-BERT [212]	brWaC (web text for Brazilian Portuguese) [2.6B]	29.8K	108.9M
pt-bio-bert [199]	Initialized with mBERT; continue on PubMed and Scielo (scholarly articles) [16.4M]	119.5K	177.9M
ro-BERT [64]	OSCAR, OPUS, Wikipedia [15.2G, 2.4B]	50.0K	124.4M
mBERT [59]	Wikipedia [72G]	119.5K	177.9M

Table 32: A comparison between baseline `mono` models and the `multi` model: mBERT. We use total file size (Gigabyte) and the total number of tokens to represent the training data size.

biomedical publications across various medical disciplines. The Spanish pharmaconer [1] dataset comprises publicly available clinical case studies, which show properties of the biomedical literature as well as clinical records, and has annotations for pharmacological substances, compounds and proteins. The English ncbi disease [61] dataset consists of PubMed abstracts annotated for disease names. The Portuguese clinpt dataset is the publicly available subset of the data collected by Lopes, Teixeira, and Gonçalo Oliveira [127], and comprises texts about neurology from a clinical journal.

prepocessing ner data We convert all annotations to BIO format. The gaps in discontinuous entities are labeled. We sentence tokenize at line breaks, and if unavailable at fullstops. We word tokenize all data at white spaces and split off numbers and special characters. If available, we use official train/dev/test splits. For bioro, we produce a random 60/20/20 split. For clinpt, we use the data from volume 2 for training and development data and test on volume 1.

A.2.3 *Financial data*

PREPROCESSING PRETRAINING DATA Sentences are tokenized using NLTK. For languages not cover by the sentence tokenizer, we split by full stops. Additionally, a split check of particular large sentences, filtering out sentences with no letters, and HTML and tags have been removed.

FINMULTICORPUS The corpus consists of PwC publications in multiple languages made publicly available on PwC websites. The publications cover a diverse range of topics that relates to the financial domain. The corpus is created by extracting text passages from publications. Table 34 describes the number of sentences and the languages that the CPT corpus cover.

FINNEWS The financial sentiment dataset is curated from financial newspapers headline tweets. The motivation was to create a Danish equivalent to FINANCIAL PHRASEBANK. The news headlines are annotated with a sentiment by 2 annotators. The annotators were screened to ensure sufficient domain and educational background. A description of *positive*, *neutral*, and *negative* was formalized before the annotation process. The dataset has an 82.125% rater agreement and a Krippendorff's alpha of .725 measured on 800 randomly sampled instances.

ONE MILLION POSTS [197] The annotated dataset includes user comments posted to an Austrian newspaper. We use the TITLE (newspaper headline) and TOPICS, i.e., 'KULTUR', 'SPORT', 'WIRTSCHAFT', 'INTERNATIONAL', 'INLAND', 'WISSENSCHAFT', 'PANORAMA', 'ETAT', 'WEB'. With the dataset, we derive two downstream tasks. The binary classification task OMP $_{\texttt{binary}}$ that deals with whether a TITLE concerns a financial TOPICS or not. Here we merge all non-financial TOPICS into one category. The multi-class classification OMP $_{\texttt{multi}}$ seeks to classify a TITLE into one of the 9 TOPICS.

A.2.4 *Adapter-based training*

Recall that the main component of a transformer model is a stack of transformer layers, each of which consists of a multi-head self-attention network and a feed-forward network, followed by layer normalization. The idea of adapter-based training [89, 166, 213] is to add a small size network (called *adapter*) into each transformer layer. Then during the training stage, only the weights of new adapters are updated while keeping the base transformer model fixed. Different options regarding where adapters are placed, and its network architecture exist. In this work, we use the bottleneck architecture proposed by Houlsby et al. [89] and put the adapters after the feed-forward network, following [166]:

Lang	PM abstracts	PM titles	M_D	M_{WIKI}
fr	54,047	681,774	735,821	872,678
es	73,704	312,169	385,873	939,452
de	31,849	814,158	846,007	831,257
it	14,031	265,272	279,303	923,548
pt	38,716	79,766	118,482	811,522
ru	43,050	576,684	619,734	908,011
ro	0	27,006	27,006	569,792
en	227,808	0	227,808	903,706
Total	483,205	2,756,829	3,240,034	6,759,966

Table 33: Number of sentences of multilingual domain-specific pre-training data for biomedical domain. Upsampling for EN was done from PM abstracts instead of Wikipedia.

$$\text{Adapter}_l\left(h_l, r_l\right) = U_l\left(\text{ReLU}\left(D_l\left(h_l\right)\right)\right) + r_l$$

where r_l is the output of the transformer's feed-forward layer and h_l is the output of the subsequent layer normalisation.

Lang	RCV2	PwC	M_D	M_{WIKI}
zh	222,308	1,466	223,774	470,111
da	72,349	192,352	264,701	465,044
nl	15,131	34,344	49,475	391,750
fr	863,911	51,500	915,411	143,427
de	1,104,603	71,382	1,175,985	0
it	138,814	22,499	161,313	467,680
ja	88,333	20,936	109,269	450,352
no	92,828	19,208	112,036	451,799
pt	57,321	35,323	92,644	439,942
ru	192,869	48,388	241,257	468,466
es	936,402	51,100	987,502	95,691
sv	132,456	25,336	157,792	467,050
en	0	346,856	346,856	444,532
tr	0	34,990	34,990	362,685
Total	3,917,325	955,680	4,873,005	5,118,529

Table 34: Number of sentences of multilingual domain-specific pretraining data for financial domain. Upsampling for EN used the TRC2 corpus instead of Wikipedia.

A.3 CHAPTER 6

A.3.1 *Annotator agreement*

The Table 35 below presents the annotator agreement on topic level. The rather high agreement across topics indicate that our annotations are of high quality.

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

Table 35: Full report of inter-annotation agreement of multi-label Cohen's κ.

A.3.2 *Label distribution*

We present the distribution of the Low-Level and High-Level topics. In Table 36, we present the distribution over the Low-Level topics. We allowed up-to 3 annotations per examples for the multi-label annotation. This produced a total of 14,230 annotation with 1.4 annotations per example on an average. In Table 37, we present the distribution over the High-Level topics.

A.3.3 *Cross-lingual transfer with fasttext embeddings*

PREPROCESSING In order to represent inputs with pre-trained fasttext embeddings, we tokenize our data according to how the fasttext

No.	Topic	Examples
1	Actuary, Pension & Insurance	502
2	Asset & Wealth Management	257
3	Accounting & Assurance	1,452
4	Banking & Financial Markets	782
5	Board, Strategy & Management	866
6	Power, Energy & Renewables	248
7	Corporate Responsibility	277
8	Media & Entertainment	255
9	Financial Crime	310
10	Government & Policy	528
11	Healthcare & Pharmaceuticals	245
12	Human Resources	1,091
13	IT Security	424
14	Governance, Controls & Compliance	501
15	M&A & Valuations	492
16	Real Estate & Construction	351
17	Retail & Consumers	354
18	SME & Family Business	226
19	Start-Up, Innovation & Entrepreneurship	277
20	Supply Chain & Transport	222
21	Tax	1,713
22	Technology	1,169
23	VAT & Customs	1,688
Total		14,230

Table 36: Overview of Low-Level tags across the 23 topics. These represent the 23 labels used in the multi-label task.

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

training data was tokenized, using Mecab[1] for Japanese, and the tokenizer from the Europarl preprocessing tools[2] [109] for the other languages.

[1] https://pypi.org/project/mecab-python3/

[2] https://www.statmt.org/europarl/

Model	Learning rate	# train epochs	# Params.
FASTTEXT$_{\text{BAG}}$	[1e-3,2.5e-3,5e-3,7.5e-3,1e-2,2.5e-2,5e-2]	50	0.1M
FASTTEXT$_{\text{LSTM}}$	[1e-3,2.5e-3,5e-3,7.5e-3,1e-2,2.5e-2,5e-2]	50	1.8M/1M/1M
sBERT	[1e-2, 3e-2, 1e-1]	[10, 30, 100]	0.6M
mBERT	[1e-5, 2e-5, 5e-5, 1e-4]	[10, 30, 100]	180M
XLM-R	[1e-5, 2e-5, 5e-5, 1e-4]	[10, 30, 100]	270M
MT5	[1e-4, 3e-4, 1e-3]	[10, 30]	300M

Table 38: The search space of two hyperparameters (learning rate and number of training epochs), as well as the number of trainable parameters for each model. The size of the hidden states in FASTTEXT$_{\text{LSTM}}$ is treated as an additional hyperparameter selected from [100,200,300,400,500], hence we report numbers of parameters for three different selected models trained on ALL/ENGLISH/HIGH RESOURCE, corresponding to models with hidden dimensionality 300/200/200, respectively. For all models, we do early stopping on the validation set with a patience of 5 and 10 for transformer-based and fasttext-based models, respectively.

EMBEDDING ALIGNMENT We map monolingual fasttext embeddings trained on Wikipedia and Common Crawl into a shared space using RCSLS, by computing pairwise mappings between source languages and English as a target language. As supervision, we rely on the training dictionaries of the MUSE dataset [48], except for Icelandic which is not covered there. For Icelandic, we follow Vulić et al. [226] in deriving a dictionary based on the Panlex database [107]: We retrieve translations for the 5000 most frequent Icelandic words derived from Opensubtitles published on Wiktionary.[3] We only keep single-word translations. As not all source words are present in Panlex, our final dictionary contains translations for 1,823 Icelandic words. With these dictionaries as supervision, we run RCSLS with default parameters for 10 epochs, and select the best mapping based on the unsupervised selection criterion.

A.3.4 *Experimental details*

For each experiment, we perform grid search to find the best combination of two hyperparameters—number of training epochs and learning rates—on the development set. Table 38 shows the search space of these two hyperparameters as well as the trainable parameters per model. The particular versions of pre-trained multilingual models can be found at:

- sBERT: https://huggingface.co/sentence-transformers/all-mpnet-base-v2
- mBERT: https://huggingface.co/bert-base-multilingual-cased
- XLM-R: https://huggingface.co/xlm-roberta-base

[3] https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Icelandic_wordlist

- MT5 https://huggingface.co/google/mt5-base

Pre-trained fasttext embeddings can be found at:

- https://fasttext.cc/docs/en/crawl-vectors.html

A.3.5 *Results of multi-class classification on* HIGH-LEVEL *topics*

Table 39 shows the evaluation results on coarse-grained categories (HIGH-LEVEL), framed as a multi-class classification problem.

Model	Training	Test		
		ALL	NO ENGLISH	LOW RESOURCE
FASTTEXT$_{BAG}$	ALL	78.1 ± 0.2	76.7 ± 0.8	70.5 ± 1.4
	ENGLISH	60.0 ± 1.0	52.2± 1.1	47.7± 1.1
	HIGH RESOURCE	73.6± 2.4	71.4 ± 2.1	52.8 ± 1.8
FASTTEXT$_{LSTM}$	ALL	83.1 ± 0.7	81.3 ± 0.8	75.9 ± 1.2
	ENGLISH	64.1± 1.5	55.7 ± 1.9	51.6 ± 2.1
	HIGH RESOURCE	80.4 ± 0.4	77.6± 0.5	60.5± 1.5
sBERT	ALL	72.4 ± 0.8	66.1 ± 1.0	55.3 ± 1.8
	ENGLISH	51.9 ± 0.5	38.4 ± 0.8	32.3 ± 0.8
	HIGH RESOURCE	72.1 ± 0.6	65.3 ± 0.7	33.0 ± 1.5
mBERT	ALL	87.4 ± 0.4	85.0 ± 0.4	79.1 ± 0.9
	ENGLISH	60.4 ± 2.4	48.4 ± 3.2	48.1 ± 2.2
	HIGH RESOURCE	82.9 ± 0.5	79.0 ± 0.7	52.3 ± 2.0
XLM-R	ALL	**89.5 ± 0.4**	**87.8 ± 0.5**	**84.0 ± 0.9**
	ENGLISH	74.9 ± 2.2	68.5 ± 2.7	67.9 ± 1.0
	HIGH RESOURCE	87.5 ± 0.7	85.3 ± 0.8	74.7 ± 1.0
MT5	ALL	83.6 ± 0.4	79.7 ± 0.5	61.3 ± 1.2
	ENGLISH	56.6 ± 0.7	42.9 ± 0.8	41.5 ± 1.3
	HIGH RESOURCE	81.1 ± 0.0	76.2 ± 0.1	43.9 ± 0.1

Table 39: Evaluation results on coarse-grained categories (HIGH-LEVEL). Results are averaged over five runs and reported by F1 micro. Multi-class classification task with 6 classes, one per example. Best results are marked with bold.

A.3.6 *Sentence length distribution*

Figure 13 shows the sentence length distribution across languages in the MULTIFIN dataset.

A.4 CHAPTER 7

A.4.1 *Inter-annotator rank agreement*

Chapter 7 presented the inter-annotator agreement for span in section 7.2, Table 24. The κ coefficients suggest a substantial agreement across all languages for the span annotation [100]. However, we proposed the rank agreement measured using RBO_{EXT} [229] as an alternative for studying differences in which words are important for humans compared to words important for the neural networks as identified by explainability methods. The main results are presented in Table 25, showing the rank agreement as measured by RBO_{EXT}. The question arises how high the inter-annotator agreement is on the trilingual dataset. Therefore, we measured the agreement between the two human rankings using RBO_{EXT} [229], which serves as a baseline for analyzing the results in Table 25. The results are given in Table 40 and show that the two annotators have a higher agreement for EN and DA than observed for IT.

Languages	Human agreement (RBO_{EXT})
DA	0.7558
EN	0.7877
IT	0.6247

Table 40: Inter-annotator rank agreement. The minimum sentence length is 3 tokens for all three languages. The maximum sentence length is 42 for both DA and EN, while 44 for IT. The average sentence length is 17 for both DA and EN, and 18 for IT [100].

Compared to the results in Table 25 and across target languages in Table 26, the agreement as measured by RBO_{EXT} between the two human annotators is higher than the agreement between machine rationales and human rationales.

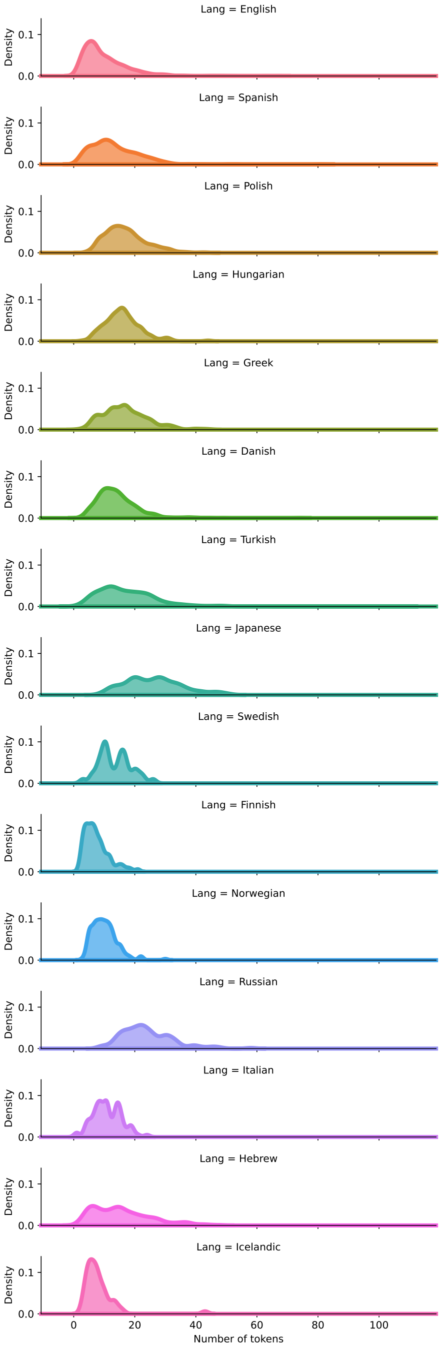


Figure 13: Sentence length distribution across different languages.

BIBLIOGRAPHY

[1] Aitor Gonzalez Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. "Pharmaconer: Pharmacological Substances, Compounds and Proteins Named Entity Recognition Track." In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, pp. 1–10.

[2] Roee Aharoni, Melvin Johnson, and Orhan Firat. "Massively Multilingual Neural Machine Translation." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3874–3884. DOI: 10.18653/v1/N19-1388. URL: https://aclanthology.org/N19-1388.

[3] Noujoud Ahbali, Xinyuan Liu, Albert Nanda, Jamie Stark, Ashit Talukder, and Rupinder Paul Khandpur. "Identifying Corporate Credit Risk Sentiments from Financial News." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022, pp. 362–370. DOI: 10.18653/v1/2022.naacl-industry.40. URL: https://aclanthology.org/2022.naacl-industry.40.

[4] Maximilian Ahrens and Michael McMahon. "Extracting Economic Signals from Central Bank Speeches." In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 93–114. DOI: 10.18653/v1/2021.econlp-1.12. URL: https://aclanthology.org/2021.econlp-1.12.

[5] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. "Publicly Available Clinical BERT Embeddings." In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 72–78. DOI: 10.18653/v1/W19-1909. URL: https://aclanthology.org/W19-1909.

[6] Mohammed Alshahrani, Fuxi Zhu, Mohammed Alghaili, Eshrag Refaee, and Mervat Bamiah. "BORSAH: An Arabic Sentiment Financial Tweets Corpus." In: *FNP 2018—Proceedings of the 1st Financial Narrative Processing Workshop@ LREC*. 2018, pp. 17–22.

[7] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. *Deep Learning Models for Multilingual Hate Speech Detection*. 2020. arXiv: 2004.06465 [cs.SI].

[8] Dogu Araci. “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.” In: *CoRR* abs/1908.10063 (2019). arXiv: 1908.10063. URL: http://arxiv.org/abs/1908.10063.

[9] Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F. Bissyandé, Jacques Klein, and Anne Goujon. “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. Ljubljana, Slovenia: Association for Computing Machinery, 2021, 260–268. ISBN: 9781450383134. DOI: 10.1145/3442442.3451375. URL: https://doi.org/10.1145/3442442.3451375.

[10] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “Translation Artifacts in Cross-lingual Transfer Learning.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7674–7684. DOI: 10.18653/v1/2020.emnlp-main.618. URL: https://aclanthology.org/2020.emnlp-main.618.

[11] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. “A Diagnostic Study of Explainability Techniques for Text Classification.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3256–3274. DOI: 10.18653/v1/2020.emnlp-main.263. URL: https://aclanthology.org/2020.emnlp-main.263.

[12] Isabelle Augenstein. “Towards Explainable Fact Checking.” In: Dr. Scient. thesis, University of Copenhagen, Faculty of Science, 2021.

[13] Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. “The FinSBD-2019 Shared Task: Sentence Boundary Detection in PDF Noisy Text in the Financial Domain.” In: *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, Aug. 2019, pp. 74–80. URL: https://aclanthology.org/W19-5512.

[14] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.” In: *PLOS ONE* 10.7 (July 2015), pp. 1–46. DOI: 10.1371/journal.pone.0130140. URL: https://doi.org/10.1371/journal.pone.0130140.

[15] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. “Interpretable Neural Predictions with Differentiable Binary Variables.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computa-

tional Linguistics, July 2019, pp. 2963–2977. DOI: 10.18653/v1/P19-1284. URL: https://aclanthology.org/P19-1284.

[16] Rachel Bawden et al. "Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages." In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 660–687. URL: https://aclanthology.org/2020.wmt-1.76.

[17] Iz Beltagy, Kyle Lo, and Arman Cohan. "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. DOI: 10.18653/v1/D19-1371. URL: https://aclanthology.org/D19-1371.

[18] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. "A Neural Probabilistic Language Model." In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf.

[19] Hampus Bengtsson and Johannes Jansson. "Using Classification Algorithms for Smart Suggestions in Accounting Systems." MA thesis. Chalmers University of Technology. Department of Computer Science and Engineering, 2015.

[20] Johan Bergdorf. "Machine Learning and Rule Induction in Invoice Processing." MA thesis. KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, 2018.

[21] Himanshu S Bhatt, Arun Rajkumar, and Shourya Roy. "Multi-Source Iterative Adaptation for Cross-Domain Classification." In: *International Joint Conferences on Artificial Intelligence (IJCAI)*. 2016, pp. 3691–3697.

[22] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

[23] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. "A large annotated corpus for learning natural language inference." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://aclanthology.org/D15-1075.

[24] Samuel R. Bowman and George Dahl. "What Will it Take to Fix Benchmarking in Natural Language Understanding?" In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 4843–4855. DOI: 10.18653/v1/2021.naacl-main.385. URL: https://aclanthology.org/2021.naacl-main.385.

[25] Steven Bramhall, Hayley Horn, Michael Tieu, and Nibhrat Lohia. "Qlime- A Quadratic Local Interpretable Model-Agnostic Explanation Approach." In: *SMU Data Science Review* 3.1 (2020), p. 4.

[26] Leo Breiman. "Random Forests." In: *Machine Learning* 45.1 (2001), pp. 5–32.

[27] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[28] Tom Brown et al. "Language Models are Few-Shot Learners." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[29] Sven Buechel, Simon Junker, Thore Schlaak, Claus Michelsen, and Udo Hahn. "A Time Series Analysis of Emotional Loading in Central Bank Statements." In: *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 16–21. DOI: 10.18653/v1/D19-5103. URL: https://aclanthology.org/D19-5103.

[30] Nadia Burkart and Marco F Huber. "A Survey on the Explainability of Supervised Machine Learning." In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[31] Thomas Burri and Fredrik von Bothmer. "The New EU Legislation on Artificial Intelligence: A Primer." In: *Available at SSRN 3831424* (2021).

[32] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. "e-SNLI: Natural Language Inference with Natural Language Explanations." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf.

[33] Joaquin Quiñonero Candela, Masashi Sugiyama, Neil D Lawrence, Anton Schwaighofer, et al., eds. *Dataset Shift in Machine Learning*. Neural Information Processing Series. MIT Press, 2009.

[34] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. "Spanish Pre-Trained BERT Model and Evaluation Data." In: *PML4DC at ICLR 2020*. 2020.

[35] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. "LEGAL-BERT: The Muppets straight out of Law School." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2898–2904. DOI: 10.18653/v1/2020.findings-emnlp.261. URL: https://aclanthology.org/2020.findings-emnlp.261.

[36] Branden Chan, Stefan Schweter, and Timo Möller. "German's Next Language Model." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6788–6796. DOI: 10.18653/v1/2020.coling-main.598. URL: https://aclanthology.org/2020.coling-main.598.

[37] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. "Numeral Attachment with Auxiliary Tasks." In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 1161–1164.

[38] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. "Numeral Understanding in Financial Tweets for Fine-Grained Crowd-Based forecasting." In: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2018, pp. 136–143.

[39] Deli Chen, Shuming Ma, Keiko Harimoto, Ruihan Bao, Qi Su, and Xu Sun. "Group, Extract and Aggregate: Summarizing a Large Amount of Finance News for Forex Movement Prediction." In: *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 41–50. DOI: 10.18653/v1/D19-5106. URL: https://aclanthology.org/D19-5106.

[40] George H Chen, Devavrat Shah, et al. "Explaining the Success of Nearest Neighbor Methods in Prediction." In: *Foundations and Trends® in Machine Learning* 10.5-6 (2018), pp. 337–588.

[41] Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. "Deep Short Text Classification with Knowledge Powered Attention." In: *AAAI*. 2019. URL: https://arxiv.org/abs/1902.08050.

[42] Yang Chen and Alan Ritter. "Model Selection for Cross-lingual Transfer." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguis-

tics, Nov. 2021, pp. 5675–5687. DOI: 10.18653/v1/2021.emnlp-main.459. URL: https://aclanthology.org/2021.emnlp-main.459.

[43] Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. "mT6: Multilingual Pretrained Text-to-Text Transformer with Translation Pairs." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1671–1683. DOI: 10.18653/v1/2021.emnlp-main.125. URL: https://aclanthology.org/2021.emnlp-main.125.

[44] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. "PaLM: Scaling Language Modeling with Pathways." In: *arXiv preprint arXiv:2204.02311* (2022).

[45] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

[46] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised Cross-lingual Representation Learning at Scale." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

[47] Alexis Conneau and Guillaume Lample. "Cross-lingual Language Model Pretraining." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

[48] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word Translation Without Parallel Data." In: *arXiv preprint arXiv:1710.04087* (2017).

[49] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. "XNLI: Evaluating Cross-lingual Sentence Representations." In: *EMNLP*. 2018. URL: https://aclanthology.org/D18-1269.

[50] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. "Semeval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News." In: Association for Computational Linguistics (ACL). 2017.

[51] Council of European Union. "Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." In: *COM/2021/206 final* (2021). URL: https://artificialintelligenceact.eu/the-act/.

[52] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. "Revisiting Transformer-based Models for Long Document Classification." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7212–7230. URL: https://aclanthology.org/2022.findings-emnlp.534.

[53] Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. "Cost-effective Selection of Pretraining Data: A Case Study of Pre-training BERT on Social Media." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1675–1681. DOI: 10.18653/v1/2020.findings-emnlp.151. URL: https://aclanthology.org/2020.findings-emnlp.151.

[54] Vinicio DeSola, Kevin Hanna, and Pri Nonis. "Finbert: Pre-trained Model on SEC Filings for Financial Natural Language Tasks." In: *University of California* (2019).

[55] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. "ERASER: A Benchmark to Evaluate Rationalized NLP Models." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408.

[56] Luciano Del Corro and Johannes Hoffart. "From Stock Prediction to Financial Relevance: Repurposing Attention Weights to Assess News Relevance Without Manual Annotations." In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 45–49. DOI: 10.18653/v1/2021.econlp-1.6. URL: https://aclanthology.org/2021.econlp-1.6.

[57] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[58] Vinicio Desola, Kevin Hanna, and Pri Nonis. "FinBERT: Pre-Trained Model on SEC Filings for Financial Natural Language Tasks." In: (Aug. 2019). DOI: 10.13140/RG.2.2.19153.89442.

[59] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "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 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[60] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. "Deep Learning for Event-Driven Stock Prediction." In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, 2327–2333. ISBN: 9781577357384.

[61] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. "NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization." In: *Journal of biomedical informatics* 47 (2014), pp. 1–10.

[62] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning." In: *arXiv preprint arXiv:1702.08608* (2017).

[63] Xin Du and Kumiko Tanaka-Ishii. "Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 3353–3363. DOI: 10.18653/v1/2020.acl-main.307. URL: https://aclanthology.org/2020.acl-main.307.

[64] Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. "The birth of Romanian BERT." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4324–4328. DOI: 10.18653/v1/2020.findings-emnlp.387. URL: https://aclanthology.org/2020.findings-emnlp.387.

[65] EUROSTAT. *EPSAS issue paper on the national approaches to harmonisation of chart of accounts*. Report EPSAS WG 17/12. 2017.

[66] Mahmoud El-Haj. "MultiLing 2019: Financial Narrative Summarisation." In: *Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*. Varna, Bulgaria: INCOMA Ltd., Sept. 2019, pp. 6–10. DOI: 10.26615/978-954-452-058-8_002. URL: https://aclanthology.org/W19-8902.

[67] Mahmoud El-Haj, Nadhem ZMANDAR, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, Aris Kosmopoulos, Blanca Carbajo-Coronado, and Antonio Moreno-Sandoval. "The Financial Narrative Summarisation Shared Task (FNS 2022)." In: *Proceedings of the The 4th Financial Narrative Processing Workshop @LREC2022*. Marseille, France: European Language Resources Association, 2022, pp. 52–61. URL: https://aclanthology.org/2022.fnp-1.7.

[68] European Commission. *Final Report of The Expert Group; Accounting Systems for Small Enterprises – Recommendations and Good Practices*. Report. 2008.

[69] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. “Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?” In: *Journal of Machine Learning Research* 15 (2014), pp. 3133–3181.

[70] Eirik Folkestad, Erlend Vollset, Marius Rise Gallala, and Jon Atle Gulla. “Why Enriching Business Transactions with Linked Open Data May Be Problematic in Classification Tasks.” In: *International Conference on Knowledge Engineering and the Semantic Web*. Springer, 2017, pp. 347–362.

[71] Olav Eirik Folkestad and Erlend Emil Nøtsund Vollset. “Automatic Classification of Bank Transactions.” MA thesis. Norwegian University of Science and Technology: Department of Computer Science, 2017.

[72] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[73] Thomas Gaillat, Manel Zarrouk, André Freitas, and Brian Davis. “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), May 2018. URL: https://aclanthology.org/L18-1423.

[74] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. “Datasheets for Datasets.” In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[75] Abbas Ghaddar and Phillippe Langlais. “SEDAR: a Large Scale French-English Financial Domain Parallel Corpus.” English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 3595–3602. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.442.

[76] Fabian Gieseke and Christian Igel. “Training Big Random Forests with Little Resources.” In: *Knowledge Discovery and Data Mining (KDD)*. ACM, 2018, pp. 1445–1454.

[77] Paul Glasserman and Harry Mamaysky. “Does Unusual News Forecast Market Stress?” In: *Journal of Financial and Quantitative Analysis* 54 (Apr. 2019), pp. 1–38. DOI: 10.1017/S0022109019000127.

[78] Yoav Goldberg. “A Primer on Neural Network Models for Natural Language Processing.” In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420.

[79] Ana Valeria González, Anna Rogers, and Anders Søgaard. "On the Interaction of Belief Bias and Explanations." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 2930–2942. DOI: 10.18653/v1/2021.findings-acl.259. URL: https://aclanthology.org/2021.findings-acl.259.

[80] Ana Valeria Gonzalez and Anders Søgaard. "The Reverse Turing Test for Evaluating Interpretability Methods on Unknown Tasks." In: *NeurIPS 2020 Workshop on Human And Model in the Loop Evaluation and Training Strategies*. 2020. URL: https://openreview.net/forum?id=y190Uu1z5Zk.

[81] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. "Learning Word Vectors for 157 Languages." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: https://aclanthology.org/L18-1550.

[82] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing." In: *arXiv:2007.15779* (2020).

[83] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740. URL: https://aclanthology.org/2020.acl-main.740.

[84] Xiaochuang Han and Jacob Eisenstein. "Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling." 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4238–4248. DOI: 10.18653/v1/D19-1433. URL: https://aclanthology.org/D19-1433.

[85] Zellig S. Harris. "Distributional Structure." In: *<i>WORD</i>* 10.2-3 (1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520. eprint: https://doi.org/10.1080/00437956.1954.11659520. URL: https://doi.org/10.1080/00437956.1954.11659520.

[86] Peter Hase and Mohit Bansal. "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5540–5552. DOI: 10.18653/v1

/2020.acl-main.491. URL: https://aclanthology.org/2020.acl-main.491.

[87] Verena Haunschmid, Ethan Manilow, and Gerhard Widmer. "audioLIME: Listenable Explanations using Source Separation." In: *arXiv preprint arXiv:2008.00582* (2020).

[88] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory." In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[89] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. "Parameter-Efficient Transfer Learning for NLP." In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2790–2799.

[90] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 328–339. DOI: 10.18653/v1/P18-1031. URL: https://aclanthology.org/P18-1031.

[91] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation." In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4411–4421. URL: http://proceedings.mlr.press/v119/hu20b.html.

[92] Allen H Huang, Amy Y Zang, and Rong Zheng. "Evidence on the Information Content of Text in Analyst Reports." In: *The Accounting Review* 89.6 (2014), pp. 2151–2180.

[93] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. "Graphlime: Local Interpretable Model Explanations for Graph Neural Networks." In: *arXiv preprint arXiv:2001.06216* (2020).

[94] James Inman. *Navigation and Nautical Astronomy, for the Use of British Seamen*. F. & J. Rivington, 1849.

[95] Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. "A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 2293–2299. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.279.

[96] Alon Jacovi and Yoav Goldberg. “Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386.

[97] Melvin Johnson et al. “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.” In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 339–351. DOI: 10.1162/tacl_a_00065. URL: https://aclanthology.org/Q17-1024.

[98] Susana Jorge, Diana Vaz de Lima, Caroline Aggestam Pontoppidan, and Giovanna Dabbicco. “The Role of Charts of Account in Public Sector Accounting.” In: *II International Congress of Public Accounting*. 2019.

[99] Rasmus Kær Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. “MultiFin: A Dataset for Multilingual Financial NLP.” In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023.

[100] Rasmus Kær Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard. “Are Multilingual Sentiment Models Equally Right for the Right Reasons?” In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 131–141. URL: https://aclanthology.org/2022.blackboxnlp-1.11.

[101] Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. “mDAPT: Multilingual Domain Adaptive Pretraining in a Single Model.” In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3404–3418. DOI: 10.18653/v1/2021.findings-emnlp.290. URL: https://aclanthology.org/2021.findings-emnlp.290.

[102] Rasmus Kær Jørgensen and Christian Igel. “Machine Learning for Financial Transaction Classification across Companies using Character-Level Word Embeddings of Text Fields.” In: *Intelligent Systems in Accounting, Finance and Management* 28.3 (2021), pp. 159–172. DOI: https://doi.org/10.1002/isaf.1500. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/isaf.1500. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.1500.

[103] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. “Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2979–2984. DOI: 10.18653/v1/D18-1330. URL: https://aclanthology.org/D18-1330.

[104] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. “FastText.zip: Compressing Text Classification Models.” In: *arXiv preprint arXiv:1612.03651* (2016).

[105] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. “Bag of Tricks for Efficient Text Classification.” In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2. 2017, 427–431.

[106] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. USA: Prentice Hall PTR, 2020.

[107] David Kamholz, Jonathan Pool, and Susan Colowick. “PanLex: Building a Resource for Panlingual Lexical Translation.” In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3145–3150. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1029_Paper.pdf.

[108] Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. “Don't Use English Dev: On the Zero-Shot Cross-Lingual Evaluation of Contextual Embeddings.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 549–554. DOI: 10.18653/v1/2020.emnlp-main.40. URL: https://aclanthology.org/2020.emnlp-main.40.

[109] Philipp Koehn. “Europarl: A Parallel Corpus for Statistical Machine Translation.” In: *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, 2005, pp. 79–86. URL: https://aclanthology.org/2005.mtsummit-papers.11.

[110] Boshko Koloski, Timen Stepišnik-Perdih, Senja Pollak, and Blaž Škrlj. “Identification of COVID-19 Related Fake News via Neural Stacking.” In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Ed. by Tanmoy Chakraborty, Kai Shu, H. Russell Bernard, Huan Liu, and Md Shad Akhtar. Cham: Springer International Publishing, 2021, pp. 177–188. ISBN: 978-3-030-73696-5.

[111] Wouter M. Kouw and Marco Loog. *An introduction to Domain Adaptation and Transfer Learning*. Tech. rep. Delft University of Technology, Department of Intelligent Systems, 2018.

[112] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. “Text Classification Algorithms: A Survey.” In: *Information* 10.4 (2019).

ISSN: 2078-2489. DOI: 10.3390/info10040150. URL: https://www.mdpi.com/2078-2489/10/4/150.

[113] Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. "Investigating Multilingual NMT Representations at Scale." 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1565–1575. DOI: 10.18653/v1/D19-1167. URL: https://aclanthology.org/D19-1167.

[114] Thibault Laugel, X. Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. "Defining Locality for Surrogates in Post-hoc Interpretablity." In: *ArXiv* abs/1806.07498 (2018).

[115] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4483–4499. DOI: 10.18653/v1/2020.emnlp-main.363. URL: https://aclanthology.org/2020.emnlp-main.363.

[116] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. "A Survey on Datasets for Fairness-Aware Machine Learning." In: *WIREs Data Mining and Knowledge Discovery* 12.3 (2022), e1452. DOI: https://doi.org/10.1002/widm.1452. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1452. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1452.

[117] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. "FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised Language Model Pre-training for French)." French. In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*. Nancy, France: ATALA et AFCP, June 2020, pp. 268–278. URL: https://aclanthology.org/2020.jeptalnrecital-taln.26.

[118] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-Based Learning Applied to Document Recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278 –2324. DOI: 10.1109/5.726791.

[119] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. "On the Importance of Text Analysis for Stock Price Prediction." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1170–1175. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1065_Paper.pdf.

[120] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." In: *Bioinformatics* 36.4 (Sept. 2019), pp. 1234–1240. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf. URL: https://doi.org/10.1093/bioinformatics/btz682.

[121] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. "RCV1: A New Benchmark Collection for Text Categorization Research." In: *Journal of Machine Learning Research* 5 (2004), 361–397. ISSN: 1532-4435.

[122] Hao-Lun Lin, Jr-Shian Wu, Yu-Shiang Huang, Ming-Feng Tsai, and Chuan-Ju Wang. "NFinBERT: A Number-Aware Language Model for Financial Disclosures." In: (2021).

[123] Zachary C. Lipton. "The Mythos of Model Interpretability." In: *Commun. ACM* 61.10 (2018), 36–43. ISSN: 0001-0782. DOI: 10.1145/3233231. URL: https://doi.org/10.1145/3233231.

[124] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *ArXiv* abs/1907.11692 (2019).

[125] Yu-Wen Liu, Liang-Chih Liu, Chuan-Ju Wang, and Ming-Feng Tsai. "RiskFinder: A Sentence-level Risk Detector for Financial Reports." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 81–85. DOI: 10.18653/v1/N18-5017. URL: https://aclanthology.org/N18-5017.

[126] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. "Finbert: A Pre-trained Financial Language Representation Model for Financial Text Mining." In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 4513–4519.

[127] Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. "Contributions to Clinical Named Entity Recognition in Portuguese." In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 223–233. DOI: 10.18653/v1/W19-5024. URL: https://aclanthology.org/W19-5024.

[128] Tim Loughran and Bill McDonald. "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." In: *The Journal of Finance* 66.1 (2011), pp. 35–65. DOI: https://doi.org/10.1111/j.1540-6261.2010.01625.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01625.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x.

[129] Tim Loughran and Bill Mcdonald. "Textual Analysis in Accounting and Finance: A Survey." In: *Journal of Accounting Research* 54.4 (2016), pp. 1187–1230. DOI: https://doi.org/10.1111/1475-679X.12123. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12123. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12123.

[130] Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. "EDGAR-CORPUS: Billions of Tokens Make The World Go Round." In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 13–18. DOI: 10.18653/v1/2021.econlp-1.2. URL: https://aclanthology.org/2021.econlp-1.2.

[131] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. "FiNER: Financial Numeric Entity Recognition for XBRL Tagging." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4419–4431. DOI: 10.18653/v1/2022.acl-long.303. URL: https://aclanthology.org/2022.acl-long.303.

[132] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From Local Explanations to Global Understanding with Explainable AI for Trees." In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[133] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777. ISBN: 9781510860964.

[134] Sholto Macpherson. *Xero's No-Code Accounting? What is It and How to Prepare For It*. www.digitalfirst.com. Accessed: June 21, 2020. 2016.

[135] Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. "Deep learning models for bankruptcy prediction using textual disclosures." In: *European Journal of Operational Research* 274.2 (2019), pp. 743–758. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2018.10.024. URL: https://www.sciencedirect.com/science/article/pii/S0377221718308774.

[136] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. "Www'18 Open Challenge: Financial Opinion Mining and Question Answering." In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 1941–1942.

[137] Pekka Malo, Ankur Sinha, Pyry Takala, Oskar Ahlgren, and Iivari Lappalainen. "Learning the Roles of Directional Expressions and Domain Concepts in Financial News Analysis." In: Dec. 2013. DOI: 10.1109/ICDMW.2013.36.

[138] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts." In: *Journal of the American Society for Information Science and Technology* (Apr. 2014). DOI: 10.1002/asi.23062.

[139] Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. "The Financial Document Causality Detection Shared Task (FinCausal 2020)." In: *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. Barcelona, Spain (Online): COLING, Dec. 2020, pp. 23–32. URL: https://aclanthology.org/2020.fnp-1.3.

[140] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. *HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection*. 2020. arXiv: 2012.10289 [cs.CL].

[141] Akira Matsui, Xiang Ren, and Emilio Ferrara. "Using Word Embedding to Reveal Monetary Policy Explanation Changes." In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 56–61. DOI: 10.18653/v1/2021.econlp-1.8. URL: https://aclanthology.org/2021.econlp-1.8.

[142] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." In: *ACM Comput. Surv.* 54.6 (2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

[143] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." In: *arXiv preprint arXiv:1301.3781* (2013).

[144] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[145] Maria Mitrofan. "Bootstrapping a Romanian Corpus for Medical Named Entity Recognition." In: *RANLP*. 2017, pp. 501–509.

[146] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[147] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. "Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition." In: *Pattern Recognition* 65 (2017), pp. 211–222. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2016.11.008. URL: https://www.sciencedirect.com/science/article/pii/S0031320316303582.

[148] Chunka Mui and William E. McCarthy. "FSA: Applying AI Techniques to the Familiarization Phase of Financial Decision Making." In: *IEEE Computer Architecture Letters* 2.03 (1987), pp. 33–41.

[149] Lee Murphy. *How algorithms will set your bookkeeping to autopilot*. https://www.theglobaltreasurer.com/2017/07/12/how-algorithms-will-set-your-bookkeeping-to-autopilot. Accessed: June 21, 2020. 2017.

[150] Nikita Nangia and Samuel R. Bowman. "Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4566–4575. DOI: 10.18653/v1/P19-1449. URL: https://aclanthology.org/P19-1449.

[151] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. "BERTweet: A pre-trained language model for English Tweets." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 9–14. DOI: 10.18653/v1/2020.emnlp-demos.2. URL: https://aclanthology.org/2020.emnlp-demos.2.

[152] Dong Nguyen. "Comparing Automatic and Human Evaluation of Local Explanations for Text Classification." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1069–1078. DOI: 10.18653/v1/N18-1097. URL: https://aclanthology.org/N18-1097.

[153] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting Good Probabilities with Supervised Learning." In: *Proceedings of the 22nd International Conference on Machine Learning*. 2005, pp. 625–632. URL: https://dl.acm.org/doi/pdf/10.1145/1102351.1102430.

[154] Eirini Ntoutsi et al. “Bias in Data-Driven Artificial Intelligence Systems – An Introductory Survey.” In: *WIREs Data Mining and Knowledge Discovery* 10.3 (2020), e1356. DOI: https://doi.org/10.1002/widm.1356. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1356. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1356.

[155] Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. “The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.” In: *Proc of BioTextMining Work*. 2014, pp. 24–30.

[156] The BAS Organisation. *The Accounting Manual 2017*. Wolters Kluwer, 2017.

[157] Daniel E O’Leary and Nils Kandelin. “ACCOUNTANT: A Domain Dependent Accounting Language Processing System.” In: *Expert Systems in Finance* (1992), pp. 253–267.

[158] Daniel E O’Leary and Toshinori Munakata. “Developing Consolidated Financial Statements using a Prototype Expert System.” In: *Applied Expert Systems* (1988), pp. 143–157.

[159] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning.” In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.

[160] Md Rizwan Parvez and Kai-Wei Chang. “Evaluating the Values of Sources in Transfer Learning.” In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 5084–5116. DOI: 10.18653/v1/2021.naacl-main.402. URL: https://aclanthology.org/2021.naacl-main.402.

[161] Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. “Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks.” In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. 2021, pp. 37–44.

[162] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation.” In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[163] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. “Semi-supervised sequence tagging with bidirectional language models.” In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1756–1765. DOI: 10.18653/v1/P17-1161. URL: https://aclanthology.org/P17-1161.

[164] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[165] Eike Petersen et al. "Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions." In: *IEEE Access* 10 (2022), pp. 58375–58418. DOI: 10.1109/ACCESS.2022.3178382.

[166] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. "AdapterHub: A Framework for Adapting Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 46–54. DOI: 10.18653/v1/2020.emnlp-demos.7. URL: https://aclanthology.org/2020.emnlp-demos.7.

[167] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617. URL: https://aclanthology.org/2020.emnlp-main.617.

[168] Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. "English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 557–575. URL: https://aclanthology.org/2020.aacl-main.56.

[169] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

[170] Barbara Plank, Dirk Hovy, and Anders Søgaard. "Linguistically debatable or just plain wrong?" In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for

Computational Linguistics, June 2014, pp. 507–511. DOI: 10.3115/v1/P14-2083. URL: https://aclanthology.org/P14-2083.

[171] John C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." In: *Advances in Large Margin Classifiers*. Ed. by Alex J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans. MIT Press, 1999, pp. 61–74.

[172] Nina Poerner, Hinrich Schütze, and Benjamin Roth. "Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 340–350. DOI: 10.18653/v1/P18-1032. URL: https://aclanthology.org/P18-1032.

[173] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. "Inexpensive Domain Adaptation of Pretrained Language Models: Case Studies on Biomedical NER and Covid-19 QA." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1482–1490. DOI: 10.18653/v1/2020.findings-emnlp.134. URL: https://aclanthology.org/2020.findings-emnlp.134.

[174] Edoardo M Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. "Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages." In: *arXiv preprint arXiv:2001.11453* (2020).

[175] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. "Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5231–5247. DOI: 10.18653/v1/2020.acl-main.467. URL: https://aclanthology.org/2020.acl-main.467.

[176] Yu Qin and Yi Yang. "What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 390–401. DOI: 10.18653/v1/P19-1038. URL: https://aclanthology.org/P19-1038.

[177] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. "Scaling Language Models: Methods, Analysis & Insights from Training Gopher." In: *arXiv preprint arXiv:2112.11446* (2021).

[178] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[179] Afshin Rahimi, Yuan Li, and Trevor Cohn. “Massively Multilingual Transfer for NER.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 151–164. DOI: 10.18653/v1/P19-1015. URL: https://aclanthology.org/P19-1015.

[180] Pranav Rajpurkar, Robin Jia, and Percy Liang. “Know What You Don’t Know: Unanswerable Questions for SQuAD.” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 784–789. DOI: 10.18653/v1/P18-2124. URL: https://aclanthology.org/P18-2124.

[181] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264.

[182] Taraka Rama, Lisa Beinborn, and Steffen Eger. “Probing Multilingual BERT for Genetic and Typological Signals.” In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1214–1228. DOI: 10.18653/v1/2020.coling-main.105. URL: https://aclanthology.org/2020.coling-main.105.

[183] Alan Ramponi and Barbara Plank. “Neural Unsupervised Domain Adaptation in NLP—A Survey.” In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6838–6855. DOI: 10.18653/v1/2020.coling-main.603. URL: https://aclanthology.org/2020.coling-main.603.

[184] Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. “Attention Can Reflect Syntactic Structure (If You Let It).” In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 3031–3045. DOI: 10.18653/v1/2021.eacl-main.264. URL: https://aclanthology.org/2021.eacl-main.264.

[185] Nils Reimers and Iryna Gurevych. "Sentence – BERT: Sentence Embeddings using Siamese BERT-Networks." In: *EMNLP-IJCNLP*. 2019. URL: https://aclanthology.org/D19-1410.pdf.

[186] Nils Reimers and Iryna Gurevych. "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. URL: https://aclanthology.org/2020.emnlp-main.365.

[187] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?" Explaining the Predictions of any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. URL: https://dl.acm.org/doi/abs/10.1145/2939672.2939778.

[188] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4902–4912. DOI: 10.18653/v1/2020.acl-main.442. URL: https://aclanthology.org/2020.acl-main.442.

[189] Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. "Is Multilingual BERT Fluent in Language Generation?" In: *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*. Turku, Finland: Linköping University Electronic Press, Sept. 2019, pp. 29–36. URL: https://aclanthology.org/W19-6204.

[190] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. "The Shapley Value in Machine Learning." In: *arXiv:2202.05594* (2022).

[191] Sebastian Ruder. *Challenges and Opportunities in NLP Benchmarking*. http://ruder.io/nlp-benchmarking. 2021.

[192] Sebastian Ruder et al. "XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10215–10245. DOI: 10.18653/v1/2021.emnlp-main.802. URL: https://aclanthology.org/2021.emnlp-main.802.

[193] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. "How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3118–

3135. DOI: 10.18653/v1/2021.acl-long.243. URL: https://aclanthology.org/2021.acl-long.243.

[194] Magnus Sahlgren, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson. "It's Basically the Same Language Anyway: the Case for a Nordic Language Model." In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 367–372. URL: https://aclanthology.org/2021.nodalida-main.39.

[195] Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. "Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment." In: *Proceedings of the Australasian Language Technology Association Workshop 2015*. Parramatta, Australia, Dec. 2015, pp. 84–90. URL: https://aclanthology.org/U15-1010.

[196] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. "Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 8415–8426. DOI: 10.18653/v1/2020.emnlp-main.676. URL: https://aclanthology.org/2020.emnlp-main.676.

[197] Dietmar Schabus, Marcin Skowron, and Martin Trapp. "One Million Posts: A Data Set of German Online Discussions." In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Tokyo, Japan, Aug. 2017, pp. 1241–1244. DOI: 10.1145/3077136.3080711.

[198] Timo Schick and Hinrich Schütze. "BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 3996–4007. DOI: 10.18653/v1/2020.acl-main.368. URL: https://aclanthology.org/2020.acl-main.368.

[199] Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. "BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition." In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, Nov. 2020, pp. 65–72. DOI: 10.18653/v1/2020.clinicalnlp-1.7. URL: https://aclanthology.org/2020.clinicalnlp-1.7.

[200] Sharath M. Shankaranarayana and Davor Runje. "ALIME: Autoencoder Based Approach for Local Interpretability." In: *Intelligent Data Engineering and Automated Learning – IDEAL 2019*. Ed. by Hujun Yin, David Camacho, Peter Tino, Antonio J. Tallón-Ballesteros, Ronaldo Menezes, and Richard Allmendinger. Cham: Springer International Publishing, 2019, pp. 454–463. ISBN: 978-3-030-33607-3.

[201] Lloyd S. Shapley. "A Value for n-Person Games." In: *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, 1953, pp. 307–318. DOI: doi:10.1515/9781400881970-018. URL: https://doi.org/10.1515/9781400881970-018.

[202] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. "Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 440–450. DOI: 10.18653/v1/P18-1041. URL: https://aclanthology.org/P18-1041.

[203] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. "BioMegatron: Larger Biomedical Domain Language Model." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4700–4706. DOI: 10.18653/v1/2020.emnlp-main.379. URL: https://aclanthology.org/2020.emnlp-main.379.

[204] Vikas Sindhwani and Prem Melville. "Document-Word Coregularization for Semi-supervised Sentiment Analysis." In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 1025–1030. DOI: 10.1109/ICDM.2008.113.

[205] Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. "BERT is Not an Interlingua and the Bias of Tokenization." In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 47–55. DOI: 10.18653/v1/D19-6106. URL: https://aclanthology.org/D19-6106.

[206] Lovisa Skeppe. "Classify Swedish Bank Transactions with Early and Late Fusion Techniques." MA thesis. KTH Royal Institute of Technology, School of Computer Science and Communication (CSC), 2014.

[207] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. "Using DeepSpeed and Megatron to Train Megatron-Turing NLG

530B, A Large-Scale Generative Language Model." In: *arXiv preprint arXiv:2201.11990* (2022).

[208] Kasper Socha. "KS@LTH at SemEval-2020 Task 12: Fine-tuning Multi- and Monolingual Transformer Models for Offensive Language Detection." In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2045–2053. DOI: 10.18653/v1/2020.semeval-1.270. URL: https://aclanthology.org/2020.semeval-1.270.

[209] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. URL: https://aclanthology.org/D13-1170.

[210] Anders Søgaard. "Explainable Natural Language Processing." In: *Synthesis Lectures on Human Language Technologies* 14.3 (2021), pp. 1–123.

[211] Iam Palatnik de Sousa, Marley Maria Bernardes Rebuzzi Vellasco, and Eduardo Costa da Silva. "Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases." In: *Sensors* 19.13 (2019). ISSN: 1424-8220. DOI: 10.3390/s19132969. URL: https://www.mdpi.com/1424-8220/19/13/2969.

[212] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. "BERTimbau: pretrained BERT models for Brazilian Portuguese." In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. 2020.

[213] Asa Cooper Stickland and Iain Murray. "BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning." In: *ICML*. 2019.

[214] Julia Strout, Ye Zhang, and Raymond Mooney. "Do Human Rationales Improve Machine Explanations?" In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 56–62. DOI: 10.18653/v1/W19-4807. URL: https://aclanthology.org/W19-4807.

[215] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://aclanthology.org/P19-1355.

[216] Shiliang Sun, Honglei Shi, and Yuanbin Wu. "A Survey of Multi-Source Domain Adaptation." In: *Information Fusion* 24 (2015), pp. 84–92.

[217] Paul C Tetlock. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." In: *The Journal of Finance* 62.3 (2007), pp. 1139–1168.

[218] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." In: *The Journal of Finance* 63.3 (2008), pp. 1437–1467.

[219] The Danish Central Business Register. *Det Centrale Virksomhedsregister (CVR)*. https://data.virk.dk. Accessed: June 21, 2020. 2019.

[220] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. "LaMDA: Language Models for Dialog Applications." In: *arXiv preprint arXiv:2201.08239* (2022).

[221] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. "FEVER: a Large-scale Dataset for Fact Extraction and VERification." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: https://aclanthology.org/N18-1074.

[222] Mesut Toğaçar, Nedim Muzoğlu, Burhan Ergen, Bekir Sıddık Binboğa Yarman, and Ahmet Mesrur Halefoğlu. "Detection of COVID-19 Findings by the Local Interpretable Model-Agnostic Explanations Method of Types-Based Activations Extracted from CNNs." In: *Biomedical Signal Processing and Control* 71 (2022), p. 103128. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103128. URL: https://www.sciencedirect.com/science/article/pii/S1746809421007254.

[223] Nicolas Turenne, Ziwei Chen, Guitao Fan, Jianlong Li, Yiwen Li, Siyuan Wang, and Jiaqi Zhou. "Mining an English-Chinese parallel Dataset of Financial News." In: *Journal of Open Humanities Data* 8 (2022).

[224] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[225] Erlend Vollset, Eirik Folkestad, Marius Rise Gallala, and Jon Atle Gulla. “Making Use of External Company Data to Improve the Classification of Bank Transactions.” In: *International Conference on Advanced Data Mining and Applications*. Springer, 2017, pp. 767–780.

[226] Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. “Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?” 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4407–4418. DOI: 10.18653/v1/D19-1449. URL: https://aclanthology.org/D19-1449.

[227] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[228] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.” In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446.

[229] William Webber, Alistair Moffat, and Justin Zobel. “A similarity Measure for Indefinite Rankings.” In: *ACM Transactions on Information Systems (TOIS)* 28.4 (2010), pp. 1–38.

[230] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. “CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data.” English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4003–4012. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.494.

[231] Sarah Wiegreffe and Ana Marasovic. “Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing.” In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 2021.

[232] Shijie Wu and Mark Dredze. “Are All Languages Created Equal in Multilingual BERT?” In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. Online: Association for Com-

putational Linguistics, July 2020, pp. 120–130. DOI: 10.18653/v1/2020.repl4nlp-1.16. URL: https://aclanthology.org/2020.repl4nlp-1.16.

[233] Xianchao Wu. "Event-Driven Learning of Systematic Behaviours in Stock Markets." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2434–2444. DOI: 10.18653/v1/2020.findings-emnlp.220. URL: https://aclanthology.org/2020.findings-emnlp.220.

[234] Xero. *How artificial intelligence and machine learning will transform accounting*. https://www.xero.com/blog/2017/02/artificial-intelligence-machine-learning-transform-accounting. Accessed: June 21, 2020. 2017.

[235] Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. "Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 978–987. DOI: 10.18653/v1/2020.coling-main.85. URL: https://aclanthology.org/2020.coling-main.85.

[236] Yumo Xu and Shay B. Cohen. "Stock Movement Prediction from Tweets and Historical Prices." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1970–1979. DOI: 10.18653/v1/P18-1183. URL: https://aclanthology.org/P18-1183.

[237] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.

[238] Linyi Yang, Eoin M Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. "Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Cassification." In: *arXiv preprint arXiv:2010.12512* (2020).

[239] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. "Finbert: A Pretrained Language Model for Financial Communications." In: *arXiv preprint arXiv:2006.08097* (2020).

[240] Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. "Do Context-Aware Translation Models Pay the Right Attention?" In: *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, Aug. 2021, pp. 788–801. DOI: 10.18653/v1/2021.acl-long.65. URL: https://aclanthology.org/2021.acl-long.65.

[241] Omar Zaidan and Jason Eisner. "Modeling Annotators: A Generative Approach to Learning from Annotator Rationales." In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 31–40. URL: https://aclanthology.org/D08-1004.

[242] Omar Zaidan, Jason Eisner, and Christine Piatko. "Using "Annotator Rationales" to Improve Machine Learning for Text Categorization." In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 260–267. URL: https://aclanthology.org/N07-1033.

[243] Omar Emilio Contreras Zaragoza. "Explainable Antibiotics Prescriptions in NLP with Transformer Models." MA thesis. Stockholm University, 2021.

[244] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1628–1639. DOI: 10.18653/v1/2020.acl-main.148. URL: https://aclanthology.org/2020.acl-main.148.

[245] Ye Zhang, Iain Marshall, and Byron C. Wallace. "Rationale-Augmented Convolutional Neural Networks for Text Classification." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 795–804. DOI: 10.18653/v1/D16-1076. URL: https://aclanthology.org/D16-1076.

[246] Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. "Low-Resource Machine Translation Using Cross-Lingual Language Model Pretraining." In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, June 2021, pp. 234–240. DOI: 10.18653/v1/2021.americasnlp-1.26. URL: https://aclanthology.org/2021.americasnlp-1.26.

[247] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. "Fine-Grained Sentiment Analysis with Faithful Attention." In: *arXiv preprint arXiv:1908.06870* (2019).

[248] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books." In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. 2015, pp. 19–27.