Universiteit Leiden

Master Computer Science

Large-scale analysis of multilingual financial news with LLMs

Name:	Milou Schamhart
Student ID:	s2182572
Date:	July 9, 2024
Specialisation:	Data Science
1st supervisor:	prof. dr. Suzan Veberne
2nd supervisor:	dr. Max Baak

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgement

I would like to express my gratitude to my first supervisor, Suzan Verberne, for her support and guidance throughout the project. Her scientific expertise and regular feedback during our biweekly meetings on the project have been invaluable in shaping this thesis. Her ability to provide simple and clear explanations of complex concepts, as well as think aloud and spar on problems along the way with me has helped me grow academically. Also, I am thankful for her support, patience, and understanding in personal matters during this project. For this strong mentorship I am truly grateful, giving me the opportunity to learn and develop as a researcher and person. Thank you, Suzan, for your patience, encouragement, and support during this project.

I would like to thank my second supervisor, Max Baak, for his experience and knowledge on my topic, and asking critical questions, and providing insightful feedback during my project. His business use-case perspective provided valuable insight, and with his help, the internship at ING for this topic was made possible. Thank you, Max, for the support and combination of business and scientific knowledge helping me forward.

I want to thank my ARIA team at ING for keeping me motivated with daily meetings and keeping me scoped at the use-case with end-user insight and technological suggestions. Also, I am thankful for the data science group at ING that I was part of, for extending my knowledge on data science related topics and use-cases within the financial industry.

Abstract

In this thesis, we conduct a comparative study on the performance of Bidirectional Encoder Representations from Transformers (BERT) models and Large Language Models (LLMs) for sentiment analysis and Named Entity Recognition (NER) tasks in the financial domain. The motivation behind this research is to evaluate the effectiveness of these models in understanding and processing language in a financial context for both English and multilingual settings.

To perform this analysis, we compile a benchmark dataset for sentiment analysis and NER of multiple open-source datasets for robust fine-tuning and evaluation. Also, we create a dataset of 100 news articles, custom for the real-world financial use-case, to assess the models' capabilities on this domain. On this data, we test several BERT models and LLMs, making an extensive analysis of the performance of the models separately, as well as comparing BERT to LLMs.

For sentiment analysis, our findings are that LLMs outperform BERT. We see that LLMs are better at generalizing to unseen data and domains, which we assume is due to their large context window and more extensive training data. For BERT, we see that the domain-specific financial model performs well within its domain, but the models have more difficulty generalizing. For the NER task, focusing on organization and location classification, BERT and LLM demonstrate similar performance levels. This indicates that the strengths of LLMs in processing broader context do not necessarily transfer to better performance of tasks that rely on recognizing specific patterns within limited context, and classification on token level, rather than text level.

These results imply that while LLMs are advantageous for tasks requiring comprehensive contextual understanding and cross-domain application, BERT remains competitive for multi-class classification on the token level, requiring pattern recognition. By this, we suggest selecting a model based on the specific requirements of the task. To consider the models' performances, extensive case-specific datasets are beneficial for extensive evaluation. Further, we suggest taking resource efficiency into account in the BERT models and LLMs.

Contents

1 Introduction

The field of Natural Language Processing (NLP) is rapidly progressing. Recent developments started with the release of the transformer architecture [VSP+17], followed up by the release of the BERT [DCLT18] models, now bringing us to the LLMs [ZZL+23, BMR+20]. These rapid advancements in NLP have opened up new possibilities for a variety of applications, including applications in the financial industry. To explore and analyze the application of LLMs in the banking industry, we collaborate with ING Group.

The ING Group is a leading European universal banking and financial services corporation headquartered in Amsterdam. ING operates in over 40 countries with its main divisions being Retail Banking, Wholesale Banking, ING Direct, and ING Insurance. Wholesale Banking is the division within ING focused on corporate clients. For Wholesale Banking clients, ING provides specialized lending, tailored corporate finance, debt, and equity market solutions, sustainable finance solutions, payments & cash management, and trade and treasury services [INGa]. Wholesale Banking Lending takes up more than half of the business division, which they say is "at the heart of most of our client relationships" [INGb]. The context of this thesis is risk management within Wholesale Banking Lending, more specifically an early warning system called ARIA (Advanced Risk Integrated Application).

An early warning system is a proactive mechanism that signals potential risks and vulnerabilities to take measures to prevent unwanted outcomes.

Early warning systems in the banking industry are applied to signal unfavorable developments for a bank in a stage where preventive measures can still be taken. Research done on this topic mainly focuses on the prediction of banking crises. In these studies, for example, dynamic Bayesian networks [DBdV16] and random forest, support vector machine, neural networks, and boosting ensembles [WZZZ21] are used. However, apart from research into banking crisis prediction and early warning systems in systemic banking risk [GMOO10, OBGO13], there has not been much research conducted on early warning systems in the banking industry.

The ARIA early warning system is a Wholesale Banking application that uses AI technologies to provide automatic insights from public news, reducing the workload of front officers and risk managers. The system currently notifies when an ING client has a high likelihood of being in a negative article related to one or more of the following topics: fraud, bankruptcy, mergers & acquisitions, sanctions, environment & climate change, and human rights. The process steps currently taken to do this analysis are explained in the pipeline in Figure 1.

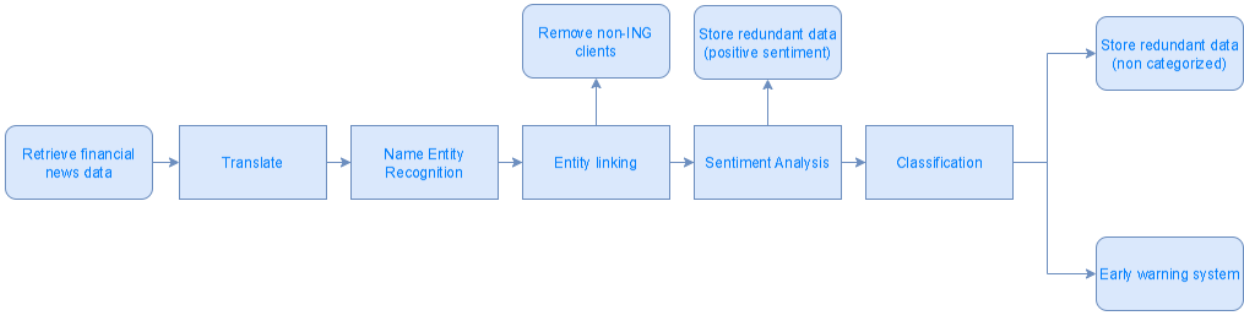

Figure 1: Pipeline from news article to early warning system

In Figure 1, the current pipeline of how the news articles are processed from raw data to a

notification in ARIA is shown. In the first step, the publicly available news articles are retrieved from Google News and the Financial Times. As the articles are from all over the world and in different languages, the next step is to translate the articles to English. After that, NER is applied to extract organizations from the article. Once the entities are extracted, these entities are linked to a database of normalized entities, to extract the organization referred to with this entity. These normalized entities are compared to a set of all organizations (ING clients) ING is interested in. If those organizations are not ING clients, the article is discarded. If the organization is an ING client, a sentiment analysis is done on the article, to see whether it has a positive or negative sentiment. If the sentiment of the article is positive, the article is stored in a backlog, as for this pipeline only articles with negative sentiment are used. So if the article's sentiment is negative, it continues to the classification step. Here, the articles are being classified based on topic. The article can be part of zero or more categories: fraud, bankruptcy, mergers & acquisitions, sanctions, environment & climate change, and human rights. If there is a match to one or more categories, a warning is created in the early warning system, for a risk manager to take action with.
For the improvement of this process, we zoom in on two processes in this pipeline, namely NER and sentiment analysis. Our approach involves the use of BERT and LLMs to process both English and multilingual news data. The current method and our improvements to this method are shown in Figure 2.

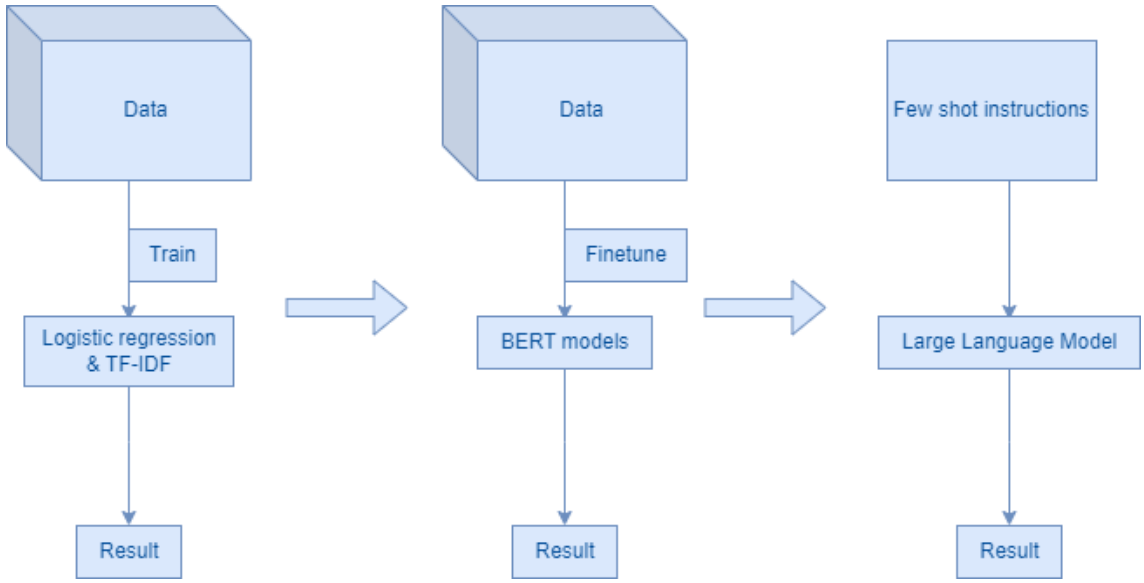

Figure 2: Transition to Large Language Models

In Figure 2, the transitional steps from ING's current method (logistic regression and TF-IDF) to the desired method (LLMs) are visualized. Currently, a method based on logistic regression and TF-IDF is trained on a lot of data, and based on that, a prediction is made [R+03, HJLS13]. TF-IDF calculates the importance of words in a document corpus, while logistic regression is a statistical method used for binary classification by predicting the probability of an event occurring. The next step in the transition would be to apply BERT models, as word-based classification methods have been outperformed for natural language processing tasks by BERT models [DCLT18]. BERT models are transformer-based architectures designed to understand the context of words by pre-training on vast amounts of text data, enabling them to generate deeply contextualized word representations. Once this has been done, we apply an approach

without using vast amounts of text data, namely generative LLMs [BMR+20]. LLMs are generative transformer models, with an extremely large amount of parameters and training data. Due to the size of these models, they can perform various NLP tasks, such as translation, sentiment analysis, NER, and classification, without fine-tuning. Using a generative model for classification tasks might seem conceptually illogical, but is made possible due to the size of these models [VSP+17, DCLT18, BMR+20]. LLMs have the purpose to understand and generate human-like language, with few-shot instructions allowing them to generalize from limited examples or instructions to perform various language-related tasks by leveraging their pre-existing knowledge and adapting to new contexts. This is a current state-of-the-art method for many NLP tasks [BMR+20].

Specifically for financial sentiment analysis, BERT models have shown to perform very well, when specifically fine-tuned for this task [Ara19]. LLMs also have been shown to perform well on financial sentiment analysis, especially when a retrieval-augmented LLMs framework is used [ZYZ+23]. When it comes to NER, fine-tuning a BERT model for supervised NER is still common practice [DCLT18, ZZ23]. However, new research bridges the gap between the two tasks of NER and LLMs: the former is a sequence labeling task, while the latter is a text-generation model. Comparable performances to fully supervised baselines are achieved [WSL+23].

Hence, our study will go deeper into comparing the performance of these different models to articles interesting for the financial domain. We will evaluate the performance of the models on two datasets, a benchmark dataset obtained from open-source data, and an ING use-case dataset created from data gathered from the in-use pipeline. Another aspect that will be taken into account in this study is the multilingualism of the data. We implement and evaluate the same techniques using the Cross-lingual Language Models (XLM). This would mean that the translation step explained in Figure 1 will no longer be needed.

To evaluate the effectiveness, we perform a comparative analysis of BERT and LLMs on English and multilingual news datasets. This involves preparing those datasets, for the multilingual aspect including low-resource languages. Low-resource languages are languages that have limited models and datasets openly available for that language [CMST16]. We will use those datasets for fine-tuning the models on labeled data and evaluating the performance. We aim to identify the most effective model by comparing these results and providing insights into the strengths and limitations of these models.

All in all, this study aims to analyze the performance of different NER and sentiment analysis models in the financial domain on English and multilingual datasets. Hence, the goal of this thesis project is to create a large-scale NLP analysis of multilingual financial news with BERT models and LLMs.

To that end, we will first analyze the performance of English and Multilingual BERT models on a corpus of English financial news as well as a multilingual corpus of financial news containing low-resource languages for sentiment analysis and NER. This leads to our first research question:

> **RQ1** How well can BERT models analyze English and multilingual financial news for sentiment analysis and NER?

The same goes for LLMs, we will analyze the performance of varying LLMs, experimenting with different prompts, on a corpus of English financial news as well as a multilingual corpus of financial news containing low-resource languages for sentiment analysis and NER. This brings us to our second research question:

RQ2 How well can LLMs analyze English and multilingual financial news for sentiment analysis and NER?

Lastly, to create a full analysis of the multilingual financial news for sentiment analysis and NER, we compare the performances of the best LLMs to that of the best BERT models to see how for ING this pipeline could be improved the most. To this end, the following research question will be addressed:

RQ3 How do the performances of BERT models and LLMs differ in sentiment analysis and NER for English and multilingual financial news?

Our addition to the current state of research will be to implement different machine learning models and compare their performance for the financial domain. Moreover, we will extensively look into the multilingual aspect, considering low-resource languages. Moreover, this thesis contributes to the field of risk management by enhancing the ARIA early warning system with recent NLP techniques. Therefore, our findings will not only benefit the banking industry but also advance the application of AI, in particular the sentiment analysis and NER tasks, in financial services. Hereby, we demonstrate the potential of machine learning in addressing real-world challenges.
With this study, we add to the existing literature by:

- Collecting open-source multilingual datasets for sentiment analysis / NER. We combine commonly used benchmark datasets for sentiment analysis and NER, to analyze model robustness and cross-domain application.
- Create a financial multilingual news articles dataset with verified sentiment / NER tags. ING bank provided previously used financial multilingual news articles, as well as domain knowledge, by which we create a case study dataset.
- Large-scale analysis of multilingual financial news with LLMs. We compare LLMs to BERT baseline in the financial domain for:
  - English NER
  - Multilingual NER
  - English sentiment analysis
  - Multilingual sentiment analysis

2 Related work

This section will cover some key papers in the field of BERT and LLM, that have caused major progress. But first, we start by explaining the NLP tasks sentiment analysis, and NER. Moreover, we include papers related to our specific use case, the application of BERT models and LLMs for sentiment analysis, and NER in the financial domain.

2.1 Sentiment analysis

Sentiment analysis is "the process of gathering and analyzing people's opinions, thoughts, and impressions regarding various topics, products, subjects, and services" [WRK22]. Typical labels for sentiment are negative, neutral, and positive when looking at opinions or general sentiment in a text. Sentiment labels can also be more domain-specific, for example when looking at reviews, labels could be 1–5 stars. Sentiment analysis is done by performing contextual mining on a text, to identify the sentiment of that text. Contextual mining can be done on different levels: aspect level, sentence level, or document level, see Figure 3. Aspect level is the smallest form of sentiment analysis, where attention is paid to the context of one or more aspects, within a sentence [WRK22]. An example of where this could be applied is opinion words in a sentence. In the sentence 'I hate Ariana Grande but I love Taylor Swift' for example, the aspects would be 'Ariana Grande', with negative sentiment, and 'Taylor Swift' with positive sentiment towards the aspect. After that, there is sentence-level sentiment analysis. This is the application of sentiment analysis separately on each sentence in the document. This is generally used when there is a wide range and mix of sentiments within one document [YC14]. When done on a document level, the sentiment analysis is performed on the whole document, giving a singular sentiment to the entire document. The application of document-level sentiment analysis has, apart from the challenge of there being multiple sentiments within one text, difficulty with cross-domain and cross-language sentiment analysis [Sau21].

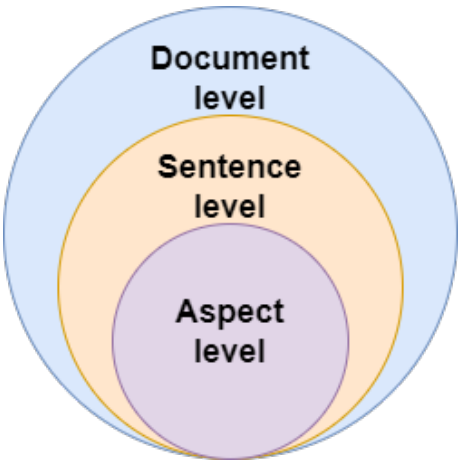

Figure 3: Level of sentiment analysis [WRK22]

2.2 Named Entity Recognition

Named entity recognition is "the task to identify mentions of rigid designators from text belonging to predefined semantic types" [LSHL20]. Commonly used predefined semantic types

are person, organization, time, location, and work of art. The process of NER consists of the first identification of the named entity (containing one or more words) in the text, often done with BERT models using IOB-labeling. The labeling with a BERT model is generally done using IOB labeling. The I stands for Inside, which is the tag given to all entities inside a labeled string of words. The O stands for Outside, this tag is given to all labels that do not belong to a named entity, meaning that it is on the Outside of the named entity. The B stands for Beginning, which indicates the beginning of a named entity. For example: 'I live in New York City', labeled as 'O, O, O, B-LOC, I-LOC, I-LOC', where LOC stands for location. By that, the named entity is classified to the predefined semantic with which it has similar attributes as other items in this semantic. It is common for NER to be applied as a pre-processing step for a variety of NLP tasks [LSHL20].

2.3 Low-resource languages

Low-resource languages are languages that "have fewer technologies and especially data sets relative to some measure of their international importance" [CMST16]. The limited availability of online resources can cause performance degradation for NLP tasks on low-resource languages. A study by Ghafoor et al. shows a performance degradation for sentiment analysis, as the translation from resource-rich languages to low-resource languages causes a polarity shift of the sentiment in the text [GID+21]. Also for NER difficulties appear with lower resource languages. Zamin et al. say that the performance of the NER task is highly domain-specific [ZOB13], making it more difficult to use cross-domain transferring techniques from resource-rich languages to lower resource languages [MCH20].

2.4 Bidirectional Encoder Representations from Transformers

Devlin et al. state that "BERT is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both left and right context in all layers" [DCLT18]. Unlike previous models that were directional (reading text from left-to-right or right-to-left) [PNI+18, RNSS18], the BERT model can learn the context of a word bidirectionally (reading text left and right of the word). This is done as the Transformer (T in BERT) encoder reads a full sequence of words simultaneously. This way, a BERT model can be a baseline model for a wide range of NLP tasks, needing only one extra layer to be fine-tuned for a specific NLP task, such as sentiment analysis or NER [DCLT18]. Fine-tuning the model for a specific task means that we provide the model with hundreds to thousands of labeled examples of what we want the model to do. In this way it can for example learn more about the meaning of words and the sentiment related to those meanings [DCLT18, Ver24].
There are different flavors of BERT models, building further on this BERT base. The distilled version of BERT is called DistilBERT [SDCW19]. This model, developed by Huggingface, is a smaller version of BERT, retaining the original architecture, but with fewer layers. This way, DistilBERT keeps 97% of BERTs language understanding, while being 60% than the original BERT model [SDCW19]. Another BERT flavor is RoBERTa, which stands for Robustly Optimized BERT Pretraining Approach [LOG+19]. This model, developed by Facebook AI, is an improved and robust version of BERT, created by an improved training method and a larger train dataset. This way, RoBERTa enhances BERT performance [LOG+19].
The application of BERT models for sentiment analysis has been studied extensively. With fitting training data, multiple studies have shown that BERT outperforms previous state-of-the-

art models for sentiment analysis [DCLT18, HBR19, AM21]. However, the BERT models are still dependent on the presence of training data and perform less well when tested cross-domain [Ara19]. Multiple studies on the application of BERT to the financial domain have shown that, with fine-tuning, a financial sentiment specified BERT can achieve results of equivalent quality. These studies focus on fine-tuning a BERT model for sentiment analysis on two financial sentiment datasets [Ara19, SSR+19, ZLZZ21]. However, they do not take multilingualism into account. There are many studies conducted on sentiment analysis, however, the majority focus on the English language [KAS+21]. Other studies focus on generalizing to one lower resource language [A+13, PVFE21, KAA+21, MSBB21], but there is a limited amount of studies going into a lexicon containing data in many different languages. This makes language coverage and domain dependence two important challenges in sentiment analysis with BERT models [Hus18, Sau21].
The NER task is also a well-known task performed by BERT models, by which this language model also outperforms previous state-of-the-art models [DCLT18]. Based on the contextual embeddings of a word sequence, the BERT model can recognize which label to give to an entity. The study by Devlin et al. shows that when fine-tuned on representative training data, BERT can accurately extract named entities out of a text [DCLT18, TKSDM03]. However, this does not directly speak for its performance cross-domain and on lower resource languages. For domains requiring a specific type of named entities, such as the biomedical domain for medicine names, fine-tuning on domain-specific labeled data is required [HP19]. The same holds for the application of multilingual NER. Features of the entities can significantly differ in different languages causing the model to have a problem with direct transfer [CKJM21]. Examples of those features for multilingualism can be the direction of writing or capitalization. Similar to the sentiment analysis, quite some research has gone into the application of NER in English or one different language [SNL19, CKJM21, LSP+22]. Also, a study performed by Wang et al. looks into extreme multilingualism, using a lexicon containing over 100 languages including lower-resource languages. Wang et al. have shown improved performance for many non-English using supervised learning compared to zero-shot, getting closer to the performance baseline of BERT applied on English datasets for NER [WMR+20].

2.5 Large Language Models

LLMs are generative language models. This refers to their ability to generate text by predicting the most likely next word after each word [RNS+18]. Also, these models are pre-trained on a large amount of data. By processing these vast amounts of data, the model can learn the probability of word sequences. Lastly, LLMs are based on transformer architecture. This is a type of model architecture, introduced in 2017 by Vaswani et al. [VSP+17], that excels at text processing and understanding word relationships. Transformers are particularly effective because they calculate the relationships between words in a text, even in long sentences, allowing the model to accurately predict subsequent words. While predicting the next word might seem trivial in itself, a transformer model can learn other language-related tasks once it is trained. To adapt a pre-trained language model for a specific task, it needs to be fine-tuned [DCLT18, LH20]. Through fine-tuning, a model can, for example, be trained to classify specific types of words in a text, as done with named entity recognition [Ver24].
The advantage of the later generation (after 2022) LLMs is that the models have many more parameters and are already pre-trained on a large amount of data, by which the models require less to no fine-tuning. This method implies using few-shot or zero-shot examples. This means

that you implement the LLM by asking it to do a specific task, without giving it any data (zero-shot) or giving it very little data (few-shot) containing examples of that task [BMR+20]. By this, LLMs apply in-context learning (ICL). With ICL, an LLM learns a task from prompt context, in which a small set of examples are included [DLD+22].
However, for an LLM to properly understand your task with little to no examples, it requires proper prompt instructions. LLMs have been shown to be prompt-prone, with the quality of the results to some extent dependent on the instructions. Literature shows there are some aspects to take into consideration for consistent and improved results from LLMs. Firstly, the chain-of-thought prompting [WWS+22], using a series of intermediate reasoning steps, has been shown to significantly improve the performance of LLMs on complex reasoning tasks. The advantages of applying this are that it is easy and effective, and adaptable to a large variety of tasks. Moreover, it is interpretable, works 'off the shelf' and increases robustness. However, downsides to applying this are that it increases the number of output tokens, which thereby can increase the production cost and decrease the clarity, as well as that with this method there is still hallucination possible [WWS+22].

2.6 Use of LLMs for classification

Supervised learning techniques for classification tasks, such as NER and sentiment analysis, have shown successful performance when enough training data is available [DCLT18]. However, for real-world use cases, the availability of the number of labeled examples is limited [WSL+23]. LLMs can be a solution for this, as they require limited to no labeled examples due to the vast size of the pre-training data. However, issues with using LLM for classification tasks are that they have been shown to suffer from a "lack of reasoning ability in addressing complex linguistic phenomena" [SLL+23], as well as there is a token limit allowed in in-context learning [SLL+23]. In this subsection, we discuss the performance of LLMs for sentiment analysis and NER, as well as the advantages and blockers that LLMs bring for these tasks.
Broekens et al. have studied the performance of LLMs on sentiment analysis, specifically the Valence, Arousal, and Dominance dimensions. They use ChatGPT zero-shot, with prompt engineering to get the results. Their findings show that an LLM is able 'to solve complex affect processing tasks emerges from language-based token prediction trained on extensive data sets' [BHV+23] and can be advanced to 'simulating, processing and analyzing human emotions'. With this, they show that with an LLM, meaningful domain-specific sentiment analysis performs very well [BHV+23]. This is an interesting indication of the high performance of sentiment analysis with sole prompting. Another study, performed by Xing, identifies the lack of knowledge on how to use generative models for domain-specific classification tasks, as well as the discriminative nature of this task as blockers for a well-performing LLM for sentiment analysis. Xing suggests a new framework with heterogeneous LLM agents. These specialized agents are instantiated using knowledge of previous errors and reasons in sentiment analysis on the aggregated agent discussions. By this, Xing shows a performance improvement, particularly when the discussions are considerable [Xin24].
Wang et al. provide insight into the performance of LLMs for NER, more specifically the LLM created by OpenAI, GPT-3, comparing it to a BERT baseline in different experimental settings. They describe that, despite the state-of-the-art performance of LLMs on many NLP tasks, for NER the performance is significantly lower than the BERT baseline. This is caused by the main purpose of the task, while LLMs are text-generation models, NER is a classification task requiring sequence labeling [WSL+23]. Wang et al. suggest bridging this gap by treating NER

as a text-generation task. Through prompt engineering they ask the LLM to reproduce the text, adding special characters around all entities. Moreover, they prompt the LLM to perform self-verification by asking itself whether the extracted entities belong to a labeled entity tag to limit hallucination [WSL+23]. Using this method on 5 widely adopted NER datasets, the performance of the LLM matches the BERT baseline. Also, Wang et al. mention that for low-resource scenarios, where there is little to no training data, the LLM significantly outperforms previous state-of-the-art supervised models [WSL+23]. A paper by Keloth et al., on NER in the biomedical domain, also underlines the shortcomings of LLMs in the effectiveness of NER. Like Wang et al., they mention this is caused by the difference in task type, as NER is a sequence labeling task, rather than a text generation task which LLMs are best at. In their research, they also transform the NER sequence labeling task into a generation task, by which they find that the performance of the LLM, in this case, LLaMA, rivals the state-of-the-art performance when it comes to multi-task, multi-domain scenarios in biomedical and health applications of NER [KHX+24].

2.7 LLMs for financial sentiment analysis

A research performed by Zhang et al. on financial sentiment analysis using retrieval-augmented LLMs shows that the direct application of LLMs for sentiment analysis causes some complications. There is a discrepancy between the sentiment analysis, which is a multi-label classification task, and the pre-training objective of LLMs, being text generation oriented. More specific to financial sentiment, Zhang et al. note that financial news is often 'devoid of sufficient context' [ZYZ+23]. To overcome these challenges, Zhang et al. suggest using the retrieval-augmented LLMs framework for financial sentiment analysis, by which they show a performance improvement compared to traditional models and LLMs like ChatGPT and LLaMA [ZYZ+23].
Another study by Ardekani et al. suggest a general financial sentiment analysis engine, "FinSentGPT" [ABB+24]. With this financial sentiment analysis specialized AI model, they take into account multilingualism. They finetune a version of ChatGPT for this purpose. FinSentGPT sentiment analysis results show to be equivalent with a state-of-the-art English-language finance sentiment model, an improvement alternative machine learning models and adds to these models by being multilingual [ABB+24].

3 Data

For this research, 3 open-source sentiment analysis datasets and 4 open-source NER datasets are merged into a benchmark dataset. These datasets are obtained from HuggingFace [hug]. Statistics and descriptions of these datasets and the final merged dataset can be found in Subsection 3.1. Moreover, there is data used from within ING. This data is a collection of news articles used for identifying financial risk, further described in Subsection 3.2.

3.1 Benchmark data

For the sentiment analysis, we combine 3 open-source Hugging Face datasets, one multilingual dataset named Multilingual Sentiment [Qia], and two financial English datasets, named financial news sentiment [Pol] and financial PhraseBank [PM]. The multilingual dataset is balanced out, meaning that when starting with a total of 282,155 samples where the vast majority is of the Chinese and Japanese languages, together covering 88.8% of the total dataset, we create a balanced dataset by having equal proportions of each language, keeping the negative, neutral, positive ratio of the original dataset. For the English financial news sentiment dataset, we keep the whole dataset and for the Financial Phrase Bank dataset, the phrases where experts agree for 75% and 100% on the annotation of the phrase are kept. A further description of the used data can be found in Table 1.

Name	Authors	#Languages	#Samples	Description
Multilingual Sentiments	Tay Yong Qiang	12	35,480	A collection of multilingual sentiments datasets grouped into 3 classes – positive, neutral, negative. The dataset is from a combination of sources, namely Twitter, Automobile platforms, Hotel reviews, Amazon, Reddit, IMDB movies, Yelp, Social media, and Online platforms. Classes in the original data sources before being grouped into the 3 mentioned classes are: 2-class positive or negative, 5-class ratings of products reviews or multiple classes of emotions, and the 3-class positive, neutral, negative [Qia]. This dataset can be found here: `https://huggingface.co/datasets/tyqiangz/multilingual-sentiments`
Financial News Sentiment	Jean-Baptiste Polle	1	1,779	Manually validated sentiment of Canadian news articles (English), grouped into 3 classes – positive, neutral, negative. Topics of these articles include: acquisition, quarterly financial release, appointment to new position, dividend, corporate update, drillings results, conference, share repurchase program, grant of stocks, and others [Pol]. This dataset can be found here: `https://huggingface.co/datasets/Jean-Baptiste/financial_news_sentiment`
Financial Phrase Bank	Pekka Malo et al.	1	5,717	An expert-annotated dataset (English) with dataset split based on agreement of 8 financial professionals making the classification (all, 75%, 66%, 50%), grouped into 3 classes – positive, neutral, negative. The dataset consists of sentences from English language financial news categorized by sentiment. All agree and 75% agree are used [PM]. This dataset can be found here: `https://huggingface.co/datasets/financial_phrasebank`

Table 1: Open-source Sentiment Analysis datasets.

The datasets are merged according to their original train and test split. For the Financial Phrase Bank dataset, there is only a split based on the agreement of experts, which we split following the current train-test data ratio (which is 85-15). The baseline characteristics are given in Table 2.

	Train English	Test English	Train Multi	Test Multi
# Samples	9,068	1,137	27,458	9,801
# English / non-English lines	9,068 / 0	1,137 / 0	9,068 / 18,390	1,137 / 8,664
# 0 / 1 / 2	1,385 / 5,050 / 2,633	302 / 446 / 389	7,515 / 11,180 / 8,763	3,202 / 3,310 / 3,289

Table 2: Characteristics of merged Sentiment Analysis datasets. 0 is negative, 1 is neutral and 2 is positive sentiment.

For the NER benchmark, we combine 4 open-source Hugging Face datasets, 2 multilingual datasets, Wikiann [RLC19] and Wikineural [TMC+21], and 2 English datasets, named CoNLLpp [TKSDM03] and Wnut17 [DNvEL17]. One of the multilingual datasets used, Wikiann, contains 176 languages. Solely 54 of those incorporated languages are kept. The languages were filtered based on the appearance of languages in the news articles retrieved by ING. For the other multilingual dataset used, Wikineural, all 9 languages are used. In Table 3, each dataset is described more thoroughly.

Name	Authors	#Languages	#Samples	Description
Wikiann	Afshin Rahimi et al.	176	1,544,500	Wikipedia articles annotated with LOC (location), PER (person), and ORG (organization) tags in the IOB2 format. The vast majority of the articles included (>90%) contain one annotation per article [RLC19]. This dataset can be found here: `https://huggingface.co/datasets/unimelb-nlp/wikiann`
Wikineural	Simone Tedeschi et al.	9	924,806	Novel technique which builds upon a multilingual lexical knowledge base (i.e., BabelNet) and transformer-based architectures (i.e., BERT) to produce high-quality annotations for multilingual NER. This method is used to automatically generate the training data for NER [TMC+21]. This dataset can be found here: `https://huggingface.co/datasets/Babelscape/wikineural`
CoNLLpp	Zihan Wang et al.	1	17,494	CoNLLpp is a corrected version of the CoNLL2003 NER dataset. The data is a collection of news wire articles from the Reuters Corpus [TKSDM03]. For CoNLLpp, 5.38% of the sentences in the test set have been manually corrected. The annotation has been done by people of the University of Antwerp [WSL+19]. This dataset can be found here: `https://huggingface.co/datasets/ZihanWangKi/conllpp`
Wnut17	Leon Derczynski et al.	1	4,681	Focuses on identifying unusual, previously-unseen entities in the context of emerging discussions. Named entities form the basis of many modern approaches to other tasks (like event clustering and summarisation), but recall on them is a real problem in noisy text - even among annotators. This drop tends to be due to novel entities and surface forms. This task will evaluate the ability to detect and classify novel, emerging, singleton-named entities in noisy text [DNvEL17]. This dataset can be found here: `https://huggingface.co/datasets/leondz/wnut_17`

Table 3: Open-source NER datasets.

The datasets are merged according to their original train and test split. After this merge, the languages are balanced under the ratio of the occurrence of each language in the news articles retrieved by ING. For this research, we are only interested in the locations and organizations

mentioned in the articles. The baseline characteristics are given in Table 4.

	Train en	Test en	Train multi	Test multi
# Samples	130,155	26,337	202,359	40,708
# English / non-English lines	130,155 / 0	26,337 / 0	130,155 / 72,204	26,337 / 14,371
# ORG / LOC tags	40,189 / 71,651	9,975 / 12,423	68,459 / 105,922	15,583 / 19,092
# Languages	1	1	54	54

Table 4: Characteristics of merged NER datasets.

3.2 Case study data

Apart from the benchmark data, we perform a financial case study on data gathered by ING. This data consists of news articles scraped from Google News of whitelisted news sources.[1] A sample of 100 articles from these scraped articles is taken and manually annotated for location, organization, and its sentiment. Due to the time intensity of this task, we were unable to extend this to a bigger sample than 100 articles. This dataset is available both fully in English (non-English articles being translated) and in a multilingual setting with all articles in their original language. For the analysis, both the title as well as the body of the article are taken into account. The baseline characteristics of these articles can be found in Table 5.

# Samples	100
# Languages	18
# EN / non-EN	54 / 46
# POS / NEU / NEG	29 / 17 / 54
# LOC / ORG	191 / 318

Table 5: Characteristics of ING test dataset.

[1]ING has a list of trusted news sources to prevent the inclusion of fake news.

4 Method

This research aims to conduct a large-scale analysis of multilingual financial data. To do this, we undertake several steps. Firstly, we gathered a vast amount of open-source data usable for large-scale training and testing of our models. Next, we labeled a dataset used by ING for the full financial analysis case study in this research. After that, we use pre-trained BERT models in zero-shot setting from HuggingFace for both sentiment analysis and NER, as well as BERT models fine-tuned specifically for these tasks on our benchmark data. Lastly, we apply sentiment analysis and NER using several LLMs to the benchmark and ING data.

4.1 Data pre-processing

To conduct the analysis, we need a robust dataset containing the desired format. To gather this, we took several steps. These differ for sentiment analysis and NER, as both have different requirements.
For sentiment analysis, we want the data to be as close to the financial domain as possible, as this influences the sentiment description. For example, for a climate news article, negative sentiment might include words like 'hazard' or 'flood', while for a tweet negative sentiment might more likely contain words like 'hate' or 'terrible'. Financial domain sentiment analysis also has specific jargon that models can learn from or be tested on, hence staying close to the financial domain is important. Another aspect is multilingualism, to train and test the models in different languages. Considering these requirements, we took several steps in pre-processing the data. For the benchmark sentiment analysis data, these steps contained:

- **Gather relevant datasets** fitting the requirements of this use case. This entails financial and multilingual sentiment analysis datasets with negative, neutral, and positive labeling.
- **Format data** to one united dataset. These datasets had different labeling, 0, 1, 2 (for negative, neutral, and positive respectively), -1, 0, 1 (for negative, neutral, and positive respectively), or the whole word (negative, neutral, positive) as a label. When merging, we make sure to use one numerical labeling system for all datasets.
- **Multilingualism** balanced based on ING language ratios. The multilingual dataset contains far more data than the English financial data. To make this more balanced, we use only part of the multilingual data of each language to have a balance (>50% English) similar to the one in the actual ING data.

Then, for sentiment analysis, we also process the ING data extracted from the currently used pipeline. These are news articles scraped from the web. We manually label these 100 articles for sentiment. When in doubt about the sentiment, the article's sentiment is double-checked by a financial expert. Due to time constraints, not all articles could be labeled by experts, creating some uncertainty in the labels. The labels are pre-processed to be in a consistent format of the labeling format of the benchmark data.
For NER, the data gathering has different requirements than for the sentiment analysis. We are only interested in two named entity types, namely locations, and organizations. These locations and organizations should not only be from the financial domain, as ING is interested in a variety of organizations from all fields on all locations. However, it wants to find locations and organizations in all their forms. For example, we want to find New York City as well as NYC or ING as well as ING Group / INGA. Hence, for the NER we try to gather a diverse

dataset containing a variety of organizations and locations. Moreover, the data for NER also has a multilingual aspect, so here the option to train and test the models on different languages is important. Considering these requirements, we took several steps pre-processing the data. For the NER benchmark data, these steps contained:

- **Gather relevant datasets** fitting the requirements of this use case. This entails English and multilingual NER datasets with a variety of organizations and locations. Also, English data containing more difficult-to-recognize entities is included.
- **Format data** to one united dataset. We use datasets that all contain IOB-labeling. However, not all IOB-labeling is consistent over the datasets. We change this to one consistent format keeping only the locations and organizations labels.
- **Multilingualism** balanced based on ING language ratios. The multilingual datasets contain far more languages than the ING data in use. Only the languages also seen in all ever-used articles by ING are kept. Also, the multilingual datasets outsize the English datasets. To make this more balanced, we use only part of the multilingual data of each language to have a balance ($>$50% English) similar to the one in the actual ING data.

Then, just like with sentiment analysis, we also process the ING data extracted from the currently used pipeline for NER. We manually label these 100 articles by organization and location. We do not apply IOB-labelling to this dataset due to time constraints. Instead, we create a list of the mentions of location and organization names mentioned in the articles.

4.2 BERT for sentiment analysis

After preparing the data, we start by applying pre-trained BERT sentiment analysis models in a zero-shot setting to our test data. We do this with 3 different models, an English financial model, an English general sentiment analysis model and a multilingual general sentiment analysis model. The financial model is a DistilRoBERTa model fine-tuned for financial sentiment analysis on the financial phrasebank dataset. With the DistilRoBERTa flavor of BERT, the model is faster than normal BERT and quality is enhanced with extra robustness. A checkpoint is used from: huggingface.co/distilroberta-finetuned-financial-news-sentiment-analysis. The general English sentiment analysis model is a RoBERTa model fine-tuned for sentiment analysis on Twitter data. This model is not distilled, but the creators have chosen to make the model extra robust in performance using RoBERTa. A checkpoint for this model is used from: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. The last, multilingual model is a DistilBERT model fine-tuned for sentiment analysis on the amazon reviews multilingual dataset. This model is the distilled version of BERT, increasing the speed significantly. For this model, a checkpoint is used from: huggingface.co/distilbert-base-multilingual-cased-sentiment.
Apart from these models, we also fine-tune a DistilBERT model for English and Multilingual financial sentiment analysis. This BERT flavor is chosen for speed and energy consumption purposes, as this model is smaller than normal BERT. For training the model for English sentiment analysis, we use only the English financial train data, so a combination of the financial Phrase Bank dataset, and the Financial News dataset. Furthermore, the learning rate is set to 2×10^{-5}, the batch size used is 16, and we use 2 train epochs. For the multilingual fine-tuned model, we use all the same fine-tuning settings, also a DistilBERT model, and the same parameter values, the only difference is that this model is fine-tuned on the English

financial data combined with multilingual data, so the a mix of English financial data and multilingual cross-domain data, making it available for multilingual use.
We apply all (both English and multilingual fine-tuned) models to the English merged test dataset and the English ING data, and apply the multilingual models to the multilingual merged test dataset and the multilingual ING data as well. All these five models provide labels negative, neutral, or positive for each data entry. We analyze these results for each sentiment category separately. Some pre-trained models give the output in a different labeling format for indicating the category. If this was not consistent with the format of our correct labels, we post-process the data to be able to calculate evaluation metrics.

4.3 BERT for named entity recognition

With the prepared NER data, we first apply an English and a multilingual pre-trained BERT model for NER from HuggingFace in a zero-shot setting to our test datasets. Firstly, we apply a large BERT model fine-tuned on the news articles in the CoNLL-2003 dataset. A checkpoint is used from: huggingface.co/bert-large-NER, which is a BERT model finetuned for NER on the CoNLL-2003 dataset. The second model we use is a DistilBERT multilingual model, fine-tuned on an aggregation of 10 high-resourced languages. The data from these languages comes from a variety of domains. Here, a checkpoint is used from: huggingface.co/distilbert-base-multilingual-cased-ner-hrl.
Moreover, we also fine-tune an English and a multilingual BERT model for NER on our benchmark train datasets. Both models are BERT models, fine-tuned on a variety of datasets containing location and organization tags in IOB labeling format. The learning rate for this fine-tuning is set to 2×10^{-5}, a batch size of 16 is used, and we only use 1 train epoch due to computation constraints in the ING environment. All parameter settings are the same for both the English and multilingual models.
We apply all these models to the English merged test dataset and the English ING data, and apply the multilingual models also to the multilingual merged test dataset and the multilingual ING data. All these models provide IOB-label formatted results for each entity in the test data set. For the benchmark test dataset, we can use these labels directly to evaluate the performance of the models for recognizing locations and organizations. For the ING test dataset however, we post-process the IOB labels back to the initial entities, to get a list of all locations and organizations for each article, as the correct labels of this test data also consist of a list of entity names, rather than IOB-labeling.

4.4 Large Language Models

In order to gather a thorough analysis of the performance of LLMs, we analyze three different models on the test data, OpenChat [WCZ+23], Llama-3 (Llama) [ml] and Gemini [AI24]. OpenChat and Llama are two openly available models. For OpenChat, a checkpoint is used from: huggingface.co/CodeNinja-1.0-OpenChat-7B-AWQ. This is a quantized version of OpenChat with 7 billion parameters. For Llama, a checkpoint is used from: huggingface.co/Meta-Llama-3-8B. Llama is developed by Meta, the mother company of Facebook. We use the 8 billion parameters version. The model was trained on a new mix of publicly available online data, however, it is not further specified what this data entails. Gemini is available through an extension in the Google Cloud Platform environment of ING. Gemini is developed by Google Deepmind and Google AI. Gemini is, like ChatGPT, available as a chatbot, but apart from

that not openly available. Hereby, also limited information on the model is publicly available. No information is shared about the training data or the number of parameters, however, it is said that the 1.5 in Gemini-1.5 stands for 1.5 trillion parameters. Either way, it is safe to assume that the number of parameters for this model exceeds the other two models by far.

4.4.1 Sentiment analysis

For sentiment analysis, we set the temperature of all models to 0.1, to give the models limited creativity, so limited hallucination of the model and getting similar results over multiple runs. However, we do not put it to 0 to introduce a slight degree of randomness. This allows minor variations while still maintaining a high degree of confidence in the output. A bit of variability can be beneficial for sentiment analysis because it allows the model to consider a wider range of interpretations and nuances in the text. This can be important when analyzing subtle differences in sentiment. To evaluate the models' robustness, as well as get optimal performances, we prompt in three steps extending the prompt each step. In the first step, we give the LLMs basic instructions on the sentiment task, providing the problem statement (PS). The first prompt can be found in the below frame.

Prompt 1: problem statement

Think like a financial risk manager at a bank.
Here is an article:

Article

The article has ended now.
What is the overall sentiment of this news article? Positive, neutral or negative?

Given your previous answer, can you provide your answer in JSON format?
For example:
'Overall sentiment': 'Positive or Neutral or Negative'

The 'Think like a financial risk manager at a bank' line helps to provide context for the LLM. Apart from that, we directly ask for the sentiment prediction. After, we ask for the output in JSON format to have a consistent output for all articles.
For the second step is an expansion of the first prompt by providing a definition (Dev.) of sentiment analysis in the context of a financial risk manager. Prompt 2 can be found in the frame below.

Prompt 2: definition

Think like a financial risk manager at a bank.
Here is an article:

Article

The article has ended now.

Sentiment for a financial risk manager is insights into public opinion towards or actions of specific companies that indicate behavior and possible trends of the company.
What is the overall sentiment of this news article? Positive, neutral or negative?

Given your previous answer, can you provide your answer in JSON format?
For example:
'Overall sentiment': 'Positive or Neutral or Negative'

The definition 'Sentiment for a financial risk manager is insights into public opinion towards or actions of specific companies that indicate behavior and possible trends of the company.' Should help the LLMs to understand how to judge the sentiment of an article and extend the context of the problem the LLM is working on.
For the third step, we apply few-shot learning. We do this by extending the prompt with 6 examples taken from the English Financial News train dataset. The examples are divided equally, so 2 of negative sentiment, 2 of neutral sentiment, and 2 of positive sentiment. This leads to the prompt found in the frame below.

Prompt 3: few-shot learning

Think like a financial risk manager at a bank.
Here is an article:

Article

The article has ended now.
Sentiment for a financial risk manager is insights into public opinion towards or actions of specific companies that indicate behavior and possible trends of the company.
What is the overall sentiment of this news article? Positive, neutral or negative?
Examples:
1. *Article*
Overall sentiment of this news article: Negative
2. *Article*
Overall sentiment of this news article: Neutral
3. *Article*
Overall sentiment of this news article: Positive
4.*Article*
Overall sentiment of this news article: Positive
6. *Article*
Overall sentiment of this news article: Negative

Given your previous answer, can you provide your answer in JSON format?
For example:
'Overall sentiment': 'Positive or Neutral or Negative'

Due to the length of these example articles, we don't provide those in the prompt 3 frame. A random set of articles was chosen. The few-shot learning is used to find if the LLMs improve when given examples of correctly identified articles related to this use case. We provide 6 examples, as literature has shown for this to be the optimal amount for LLMs [WWS+22].

We apply these prompts for all models on the ING multilingual and English test datasets. Due to computational restrictions, we only apply these prompts to the English and multilingual benchmark test data once, using Gemini. Moreover, OpenChat and Llama are not providing the result in the correct JSON format, hence we post-process those results looking for 'Negative', 'Neutral', or 'Positive' in the output, as they do consistently provide the prediction somewhere in the text output.

4.4.2 Named entity recognition

For NER, we set the temperature to 0, as we do not want the models to hallucinate and only to extract exact words from the articles, as well as make the results reproducible over multiple runs. Just like with sentiment analysis, for NER we also build up the prompt in 3 steps, starting with the problem statement. This prompt can be found in the frame below.

Prompt 1: problem statement

Here is an article:

Article

The article has ended now.
What are the organizations and locations mentioned in the article?
We are interested in all occurrences of organizations and locations in the article.

Provide your answer in JSON format.
JSON format:
"organizations": ["organization1", "organization2"],
"locations": ["locations 1", "locations 2"]

This prompt requests organizations and locations, stating the problem. We add to this the line 'We are interested in all occurrences of organizations and locations in the article' to make sure the model does not only provide each company found in an article once.
For the second step, we add some more context and explanation of this problem. The second prompt is shown in the frame below.

Prompt 2: definition

We're performing named entity recognition. I will give you an article, please identify all organizations and locations.
We are interested in all occurrences of organizations and locations in the article. Please make sure to extract the exact occurrence of the entities from the article.
Pay attention to the fact that the organizations and locations in the article can consist of multiple words.
Here is an article:

Article

The article has ended now.

What are the organizations and locations mentioned in the article?

Provide your answer in JSON format.
JSON format:
"organizations": ["organization1", "organization2"],
"locations": ["locations 1", "locations 2"]

For this second prompt, we add the task name, 'named entity recognition', as well as some additional information to pay attention to in recognizing the entities. This way we extend the problem statement by further definition of the task. The way of elaborating the task and what to pay attention is the same as instructions you could give to a human for this task.
For the third step, we again apply few-shot learning, providing six lines from the English benchmark train dataset with the correctly annotated locations and organizations. This leads to the prompt in the frame below.

Prompt 3: few-shot learning

We're performing named entity recognition. I will give you an article, please identify all organizations and locations.
We are interested in all occurrences of organizations and locations in the article. Please make sure to extract the exact occurrence of the entities from the article. Pay attention to the fact that the organizations and locations in the article can consist of multiple words.
Here is an article:

Article

The article has ended now.
What are the organizations and locations mentioned in the article?
Here are some articles as examples, with the correct organizations / Locations:
1. *Article*
"organizations": ["Reuters"]
"Locations": ["Chinese", "Taiwan", "Taiwan"]
2. *Article*
"organizations": ["Lloyds Shipping"]
"Locations": ["Syria"]
3. *Article*
"organizations": ["Opel AG", "General Motors"]
"Locations":[]
4. *Article* "organizations": ["Lloyds Shipping Intelligence Service", "Lloyds Shipping"]
"Locations": ["Syria", "LONDON", "US", "UK", "JP", "FR"]
5. *Article*
"organizations": ["Den Norske Stats Oljeselskap AS", "Statoil"]
"Locations": ["Heidrun", "mid-Norway"]
6. *Article*
"organizations": ["NCB"]
"Locations": ["Finland"]

Provide your answer in JSON format.

JSON format:
"organizations": ["organization1", "organization2"],
"locations": ["locations 1", "locations 2"]

Again, due to the length of the example articles, we do not provide the full article text in the prompt 3 frame, but these are articles randomly selected from the English benchmark train dataset. We provide 6 examples again [WWS+22], to help the LLMs understand the desired output.
We apply these prompts for all these models on the ING multilingual and English test datasets. Due to computational restrictions, we only apply these prompts to the English benchmark test data once, using Gemini. Moreover, OpenChat and Llama are not providing the result in the correct JSON format, so several post-processing steps are required. Due to time restrictions, we only post-process Llama to have the final result. To post-process the Llama output, we put exception rules for different output formats often seen to transfer to a singular output format, and after that manually go over all output to add output provided in formats not included in the exceptions. Lastly, to evaluate the output of Gemini for benchmark test data, we process the benchmark test data from IOB labels to a list of organization and location entities for each data point.

4.5 Evaluation metrics

To evaluate the predictive ability of the sentiment analysis and NER models, we calculate the F1-score. The F1-score is the harmonic mean of the precision and recall, calculated with the following formula: $F1 = 2 * \frac{precision*recall}{precision+recall}$. Precision in its turn evaluates the fraction of correctly classified data points among all data points classified as positives. Precision is calculated using the following formula: $precision = \frac{TP}{TP+FP}$. TP stands for true positives, so all data points are correctly classified as positives. For sentiment analysis, this is a correct sentiment of the whole text, and for NER this is a correctly labeled entity, so a full location or organization. FP stands for false positives, so these are all the data points identified as positives, while not belonging to this category. The recall is the fraction of true positives amongst the total number of elements that belong to the positive class. The formula used to calculate the recall is: $recall = \frac{TP}{TP+FN}$. In this formula, FN stands for false negative, so all the data points belong to the positive class but are not labeled as such.
For sentiment analysis, the F1-score is calculated per sentiment category, so for negative, neutral, and positive. This means that, for the negative class, for example, we check in a binary manner whether the data point is classified as such and whether the correct label was negative. So to get all true positive predictions, we take all articles labeled as negative by the model, as well as their correct label. For false positives, we take all articles labeled by the model as negative, while their correct label should not be negative (either neutral or positive). For false negatives, we take all articles whose correct label is negative, while the model labels it differently (either neutral or positive). The same method goes for neutral and positive.
For NER, the F1-score is calculated separately for locations and organizations. Here, for example with locations, an entity is classified as a true positive when the label predicted by the model is the location (B-LOC and all I-LOCs match in the case of IOB labeling), as well as the correct label of that entity. We evaluate on the complete entity level, not the token level. The full entity (location or organization) predicted needs to match the full entity in the labeled data. For false positive, an entity is labeled as location, while the actual label is something else (organization or outside of named entity categories). A false negative in this case would

be an entity being classified as an organization or as outside of named entity categories, while the correct label is location.

5 Results

In this section, the results obtained by experimenting with different BERT models and LLMs for sentiment analysis and NER with benchmark data and ING data are discussed. The experimental analysis is split up in the same structure as the research questions, starting with BERT, then LLMs, and lastly a comparison of the best BERT and LLM results.

5.1 BERT for sentiment analysis and NER

To answer the first research question, *How well can BERT models analyze English and multilingual financial news for sentiment analysis and NER?*, we look into the performance of BERT models of the merged datasets, as well as on the case study data from ING.

5.1.1 Sentiment analysis on benchmark data

The results of five BERT models for sentiment analysis, with different fine-tuning, on the English Financial merged benchmark data are shown in Table 6. For this table, as well as for Tables 7 – 12 holds that the 'Base model' refers to the version of BERT used, the 'Domain' refers to domain of the data used for fine-tuning this model, 'Fine-tuned' refers to whether the model is fine-tuned on the English / multilingual benchmark data (Yes), or a checkpoint from HuggingFace is used, so a pre-trained model is exposed to the new test data in zero-shot setting (No).

Base model	Language	Domain	Fine-tuned	F1 negative	F1 neutral	F1 positive	Avg. F1
DistilRoBERTa	English	Financial news	No	58.8%	**91.0%**	**87.0%**	**78.9%**
RoBERTa	English	Twitter data	No	50.2%	80.4%	46.0%	58.9%
DistilBERT	Multilingual	Amazon reviews	No	33.2%	7.8%	18.7%	19.9%
DistilBERT	English	Financial data	Yes	52.6%	86.3%	82.3%	73.7%
DistilBERT	Multilingual	Benchmark data	Yes	**75.0%**	74.3%	62.3%	70.5%

Table 6: Information on the sentiment analysis model, how it has been fine-tuned and its performance on the merged English Financial benchmark data.

Financial domain models Looking at Table 6, we find that when applying sentiment analysis to an English Financial dataset, models fine-tuned specifically on this domain perform best. Overall, DistilRoBERTa externally fine-tuned on financial news, used with a checkpoint from Huggingface, performs best with an average F1 of 78.9%, closely followed by the model we fine-tuned for English financial sentiment, the DistilBERT model with financial data as a domain with an average F1 score of 73.7%. However, the performance of the models in categorizing negative sentiment is notably lower. This can be related to the ratio of each sentiment category in the fine-tuning data of these models. In this data, the vast majority of financial news is neutral (>50%), while there is only slightly over 10% of negatively labeled financial news. This is reflected in the performance of these models on the financial test set. Moreover, looking at the precision and recall, we find that for both models for negative classification, the recall is much lower than the precision. The relatively high precision indicates a low false positives rate, meaning that the models do not label many articles as negative that are not. The lower F1 is caused by the relatively low recall, indicating a high false negative

rate. This means that the models miss a lot of articles that should have been classified as negative. This is also in line with what can be expected when the fine-tuning data has much fewer indices of one class. While there are differences in the precision and recall of the neutral and positive classes, these values are much closer together relatively, compared to the negative class.

Twitter domain model Looking at the second model in Table 6, the RoBERTa model externally fine-tuned on Twitter data, used with a checkpoint from Huggingface, we see that it performs substantially less well, especially for identifying positive sentiment. The performance of BERT models is specifically dependent on their fine-tuning and, as discussed in Section 2, have trouble working in cross-domain settings. The Twitter data significantly differs from the financial data the model is tested on. This model performs much better on neutral classification compared to negative and positive. Hence, we can assume there is also a bias towards neutral, considering this is the most common label in the fine-tuning data.

Multilingual models The performance decrease with cross-domain application for BERT models is shown even stronger when we look at the performance of DistilBERT, externally fine-tuned on multilingual Amazon reviews, used with a checkpoint from Huggingface. Like the Twitter data model, the Amazon news reviews on which this model had been fine-tuned externally is a very different domain, the financial news, especially considering that the Amazon reviews are initially ranked with 1 to 5 stars, converted to negative, neutral, and positive categories. It can be assumed, as is shown in the fine-tuning data of this model, that people are most likely to leave a review when they are unhappy about something, and least likely to leave a review when they feel neutral towards something. This is reflected in the result performance as well. However low, the negative classification done by this model is substantially better than the neutral classification.
Moreover, an extra difficulty dimension is added by the multilingual aspect. Apart from being domain-specific, the Amazon review data also contains multiple languages, on which the model is trained, which will be received as noise for the model as it is tested here on only English data.
Lastly, there is the multilingual model fine-tuned for this use case, on the whole of the gathered and merged sentiment analysis benchmark data. This model performs well, especially when looking at the classification of negative sentiment. For English, this model has learned the structure of the English financial data. However, unlike the other English financial models, the fine-tuned data of this model is more balanced over the negative, neutral, and positive classes.

Base model	Domain	Fine-tuned	F1 negative	F1 neutral	F1 positive	Avg. F1
DistilBERT	Amazon reviews	No	52.2%	14.8%	55.3%	40.8%
DistilBERT	Benchmark data	Yes	**60.5%**	**65.9%**	**60.9%**	**62.4%**

Table 7: Information on the multilingual models, how they have been fine-tuned, and their performance on the merged multilingual benchmark data.

In Table 7, results of the multilingual models (fine-tuned on the Benchmark data, as well as a pre-trained model used in zero-shot setting from HuggingFace) on the multilingual test data are shown. As explained in section 3, the benchmark multilingual test dataset does not solely contain financial news due to the lack of multilingual financial news. Hence, it is a combination

of English financial news, as well as cross-domain multilingual news. In Table 7 we find that the model fine-tuned on the benchmark data outperforms the multilingual sentiment model from HuggingFace, with an average F1 of 40.8% and 62.4%, respectively. This can be explained by the fact that, as mentioned before, the Amazon review data does not generalize well to this use-case. However, in this scenario, it does overcome the multilingual, cross-domain data better than with only English financial data as seen in Table 6.
When a BERT model is trained and tested on data of similar structure, as is the case for the benchmark multilingual train and test data, somewhat higher F1 scores can be expected. However, we speculate that the relative lack of increase is due to the variety of domains included in this data, making it harder for the model to find a recurring structure. This can explain the increased difficulty for this model, despite being multilingual trained, in this setting compared to the setting of English financial data in, as seen in Table 6.

5.1.2 Sentiment analysis on ING data

Now we discuss the results of applying those same models to our ING use-case data. However, an important note with all these results is that it is hard to fully interpret the results, as the ING test dataset only consists of 100 data points. Due to this, we can question the representativeness of this dataset compared to the whole of the incoming news article stream. Moreover, proper analysis of the errors in the models is harder as a small sample size like this chance is of higher influence on the result.

Base model	Language	Domain	Fine-tuned	F1 negative	F1 neutral	F1 positive	Avg. F1
DistilRoBERTa	English	Financial news	No	48%	28%	40%	39%
RoBERTa	English	Twitter data	No	47%	**33%**	32%	37%
DistilBERT	Multilingual	Amazon reviews	No	76%	0%	59%	45%
DistilBERT	English	Financial data	Yes	7%	21%	43%	24%
DistilBERT	Multilingual	Benchmark data	Yes	**87%**	0%	**73%**	**53%**

Table 8: Information on the model, how it has been fine-tuned, and its performance on the English ING data.

In Table 8, we see the performance of all models on the English ING news articles. This means that all non-English articles are translated into English, so we have 100 ING news articles to test on. Generally, the performance of most of these models on the ING data is lower when compared to the English benchmark data in Table 6.

Domain change To explain this trend, we dive deeper into the ING data. A qualitative analysis shows that these news articles are recent and of a short time frame, containing a bias in what news tabloids printed in that period. It is notable that with that relatively many news articles are related to the war between Russia and Ukraine, causing these articles to be in a specific domain. This dimension increases the difficulties for the models.

English fine-tuned models Looking at the first model, DistilRoBERTa, externally fine-tuned on financial news data, used with a checkpoint from Huggingface, we see that its performance now lies close to the performance of the second model, RoBERTa externally fine-tuned on Twitter data, also used with a checkpoint from Huggingface, with an average

F1 of 39% and 37%, respectively. The fact that these models now have similar performances indicates that the ING data has a similar distance from both domains. The distance of the DistilBERT model that has been fine-tuned externally on the financial benchmark data seems to be bigger to the ING data than the other two models.
Moreover, there is a substantial performance decrease for the financially fine-tuned models compared to the result achieved on the benchmark data. The DistilRoBERTa externally fine-tuned on financial news data goes from a average F1 of 78.9% on the benchmark data to an average F1 of only 39% on the ING data. The DistilBERT model that is fine-tuned on the financial benchmark data goes from an average F1 of 73.7% on the benchmark data to an average F1 of only 24% on the ING data.

Multilingual fine-tuned models That said, we do see an improvement in the performance of the multilingual DistilBERT model externally fine-tuned on the Amazon review data, used with a checkpoint from Huggingface. Despite the small size of this dataset, it shows substantial improvement compared to its performance on the financial benchmark data, going from an average F1 of 19.9% on the benchmark data to an average F1 of 45% on the ING test data. This indicates that the distance between the Amazon review data and the ING data is less compared to the financial benchmark data. Apart from neutral, the DistilBERT multilingual model fine-tuned on the benchmark data outperforms the other models for the negative and positive sentiment, resulting in the highest average F1 of all models, being 53%. A possible clarification for this could be that the multilingual benchmark data contains data from different domains and thus generalizes better to unseen domains than the other models. However, this is speculative.
However, like the result on the financial benchmark data of this model, identifying negative sentiment is easier for this model than positive, and especially neutral. The F1 for neutral is 0%. It never gets neutral correct. Taking a closer look at the predictions in this result, we find that the model only predicts neutral 2 times, once when the actual label is negative and once when the actual label is positive. We also see an F1 of 0% for neutral for the multilingual model fine-tuned on our benchmark data. Looking at the data for this model, we see that it predicts neutral only 6 times, and again none of these times are correct.

Base model	Domain	Fine-tuned	F1 negative	F1 neutral	F1 positive	Avg. F1
DistilBERT	Amazon reviews	No	**71%**	**9%**	43%	41%
DistilBERT	Benchmark data	Yes	69%	7%	**61%**	**46%**

Table 9: Information on the multilingual models, how they have been fine-tuned, and their performance on the multilingual ING data.

In Table 9, we find the results of applying multilingual models to the 100 data points ING news articles. Here, these 100 news articles are in their original language. The performance of both multilingual models is quite similar, though the model fine-tuned on the benchmark data overall outperforms the model externally fine-tuned on the Amazon review data with an average F1 of 46% compared to an average F1 of 41%, respectively. Also, this result is mostly in line with previously seen results. Again, it shows a much lower F1-score for the neutral sentiment, and compared to the performance on English ING data, there is a slight performance decrease. This can be because there are 12 languages the model is trained on, while the ING dataset contains 18 languages. Hence, not all languages are covered in the train datasets of these models.

5.1.3 NER on benchmark data

We apply two different models fine-tuned for NER, with a checkpoint from HuggingFace in a zero-shot setting. The performance of these models on the benchmark data can be found in Table 10.

Base model	Language	Domain	F1 location	F1 organization	Avg. F1
BERT	English	News articles	37.0%	**19.0%**	**28.0%**
DistilBERT	Multilingual	Cross domain*	37.0%	18.8%	27.9%
DistilBERT	Multilingual	Cross domain*	**38.2%**	15.7%	27.0%

Table 10: Information on the models from HuggingFace used in a zero-shot setting, how it has been fine-tuned, and its performance on the merged benchmark data. The first two rows are on the English benchmark data, the last, separate row is on the multilingual benchmark data. * In most languages it's news articles, in some languages it's other data types.

In Table 10, we find all both models have similar performance. The English BERT model performs best, with an average F1 of 28.0%. All models are better at finding locations compared to organizations. There is no substantial difference in performance between English and multilingual. Overall, we see that the performance of the models is quite low. This can be related to the fact that we use a merge of different open-source datasets to test our models. A model is usually adjusted to the structure of the training data. In this case, the models get data of locations and organizations in different structures due to the merge, hence it is more difficult for the model to learn this and transfer its knowledge.

5.1.4 NER on ING data

Now we discuss the results of applying those same models as before, as well as a multilingual and English model fine-tuned on the benchmark train dataset to the ING data. However, an important note with all these results is that it is hard to fully clarify the results, as the ING test dataset only consists of 100 data points. Due to this, we can question the representativeness of this dataset for the whole of the incoming news article stream. Moreover, proper analysis of the errors in the models is harder as a small sample size like this chance is of higher influence on the result.

Base model	Language	Domain	Fine-tuned	F1 location	F1 organization	Avg. F1
BERT	English	News articles	No	30%	31%	31%
DistilBERT	Multilingual	Cross domain*	No	**31%**	31%	31%
BERT	English	Benchmark data	Yes	24%	54%	39%
BERT	Multilingual	Benchmark data	Yes	29%	**55%**	**42%**

Table 11: Information on the model, how it has been fine-tuned, and its performance on the English ING data. * In most languages it's news articles, in some languages it's other data types.

In Table 11, we find that for the ING data, all models seem to perform equal or better at the classification of organizations compared to locations. The main difference compared to

the benchmark data is the increase in performance of classifying organizations. This can be attributed to the decrease of 'difficult' organizations, which are included more in the benchmark data with for example the Wnut17 dataset. Overall, the BERT model fine-tuned on the multilingual benchmark data performs best, with an average F1 of 42%. However, again like the benchmark data, the results are quite low. A possible reason for this is the domain-specific application.
Also, diving deeper into the data, we can see there is a bias in the data as all 100 news articles are from a relatively short time frame. Hereby, we notice that the topic relevant in the world at that moment is seen in news articles way more than other topics, influencing the organizations and locations mentioned. For example, with the current war between Russia and Ukraine, these countries are mentioned regularly in the articles, while many other countries are never mentioned. Also, organizations related to this are mentioned more often, such as Sberbank. As with a small dataset, there cannot be full inclusion and variety. However, we speculate this bias in these 100 news articles still to be of influence on the result.

Base model	Domain	Fine-tuned	F1 location	F1 organization	Avg. F1
DistilBERT	Cross domain*	No	**21%**	27%	24%
BERT	Benchmark data	Yes	13%	**41%**	**27%**

Table 12: Information on the multilingual models, how they have been fine-tuned, and their performance on the multilingual ING data. * In most languages it's news articles, in some languages it's other data types.

In Table 12, we find the results of applying the multilingual trained models to the multilingual ING data. On the multilingual data, the performance of these models decreases compared to the English data performance seen in Table 11. Again, both models perform better for classification for organizations than locations, but for the DistilBERT model from HuggingFace used in a zero-shot setting, the difference in performance is not substantial. The multilingual data presents an extra challenge as the models are trained in these languages but the results must be converted back to English, as only English labels are available. This makes it sensitive to translation errors, possibly being of influence on the results.

5.2 LLMs for sentiment analysis and NER

In order to answer our second research question, *How well can LLMs analyze English and multilingual financial news for sentiment analysis and NER?*, we now discuss the results of applying LLMs to our benchmark data, but due to limitations, mostly to our ING dataset.

5.2.1 Gemini on benchmark data

Due to computational limitations, it was not possible to run OpenChat or Llama on the benchmark data using ING resources. However, we were able to perform sentiment analysis and NER using Gemini.
In Table 13 we see the performance of Gemini on the English and multilingual benchmark data. We find in Table 13 that Gemini on the benchmark data performs very well. There is a trend showing that the model performs less well on positive categorisation, compared to the other sentiment categories. Moreover, we find that the model performs better on the English data compared to the multilingual data. This is likely to be due to the increased difficulty of

Input	English			Multilingual		
	F1 negative	F1 neutral	F1 positive	F1 negative	F1 neutral	F1 positive
Prompt 1: PS	**88.3%**	84.5%	75.6%	**76.4%**	67.6%	**67.5%**
Prompt 2: Def.	87.9%	**88.6%**	**76.4%**	74.8%	69.1%	59.5%
Prompt 3: Few-shot	88.2%	88.3%	69.8%	69.8%	**69.3%**	45.8%

Table 13: Results of applying Gemini for sentiment analysis on benchmark data.

multilingual settings, as it contains many languages, while the vast majority of the Gemini train data is highly likely to be English. Also, looking at the best performance for each sentiment category in bold in Table 13, we find that the results are similar between different prompts. This indicates that the Gemini model is quite robust in terms of prompt influence as there are no substantial differences between the results for all prompts. This can likely be attributed to the vast amount of parameters of the model, as well as the large size of the training data.

Input	F1 location	F1 organization
Prompt 1: PS	18.5%	14.9%
Prompt 2: Def.	**27.8%**	**26.4%**
Prompt 3: Few-shot	22.8%	19.2%

Table 14: Results of applying Gemini for NER on English benchmark data.

In Table 14, we see the result of applying Gemini to the English benchmark data. While the performance of Gemini for sentiment analysis seen in Table 13 worked very well, its performance on the test dataset for NER is much lower. The best-performing models get it right in approximately one-fourth of the time. Taking a look at the precision and recall of these results, we find that the recall is very low, as the model predicts a lot of false positives. There are over four times more tokens labeled as location and organization than there are in the correct labels. This indicates that it is quite hard for the model to correctly label tokenized data for NER. Also, we see that the second prompt, with a more extensive definition of the task, performs best, without using few-shot learning.

5.2.2 Sentiment analysis

We apply OpenChat, Llama, and Gemini to the ING data for sentiment analysis. These results are shown in Table 15. In this table, the results are divided based on the prompts per model mentioned in Section 4.

We can see in Table 15 that from these three LLMs OpenChat performs the least well. Using the prompt engineering, we do see a slight increase in overall performance in the model when more explanation is provided. OpenChat and Llama are both much smaller models than Gemini, however, OpenChat still performs substantially worse than Llama. This could be because, apart from being the smallest model taken into consideration here, we use a quantized version of the model, which is beneficial for size reduction, increasing speed, and lower power consumption. However, this can also lead to a reduction in performance. We speculate this to explain the performance difference between OpenChat and Llama.

Apart from starting out better than OpenChat, for Llama we also see performance increase when extending the prompt. While for English, the extension of the prompt with an elaborate definition does not increase Llama performance, when we compare the result provided by

Model	Input	English			Multilingual		
		F1 negative	F1 neutral	F1 positive	F1 negative	F1 neutral	F1 positive
OpenChat	Prompt 1: PS	64%	11%	0%	62%	10%	0%
	Prompt 2: Def.	53%	27%	6%	56%	24%	12%
	Prompt 3: Few-shot	53%	24%	22%	60%	31%	17%
Llama	Prompt 1: PS	86%	35%	59%	56%	31%	55%
	Prompt 2: Dev.	52%	26%	62%	64%	41%	55%
	Prompt 3: Few-shot	87%	39%	**83%**	89%	33%	**81%**
Gemini	Prompt 1: PS	**89%**	40%	80%	**90%**	**54%**	75%
	Prompt 2: Dev.	86%	**46%**	**83%**	**90%**	46%	75%
	Prompt 3: Few-shot	84%	39%	67%	84%	53%	71%

Table 15: Results of applying LLMs for sentiment analysis on ING news data.

the initial prompt to the result of the 3rd, final prompt, we do see a substantial increase in performance. For the multilingual setting, the performance increases steadily with each extension of the prompt. With these final results, the results of the Llama model using the 3rd prompt are competitive with the Gemini results for both English and multilingual settings.
Gemini is the best-performing model. What we can clearly notice, is that Gemini is much less prone to prompt engineering. As shown in the multilingual setting, the model outperforms the other models and the other, more extensive prompts already in the 1st prompt. For the English test data, the second prompt provides the best performance, however, this is not a very substantial increase compared to the first prompt. This is due to the size of this model. Gemini has far more parameters than the other models, as well as a bigger training dataset, and this makes it good at tasks it is not specifically familiar with. Lastly, we note that the 3rd prompt for both English and multilingual settings is showing a decrease in performance. We speculate that the model might get confused by the extra examples, which it does not specifically learn more from.
Overall, Gemini performs best, but when applying prompt engineering, Llama reaches competitive performance. Taking into consideration that these models do not perform the same results twice, these differences are minimal. Lastly, it is notable that for all models, there is no substantial difference in the performance of the English and Multilingual data. This shows that these larger models are not language-prone.

5.2.3 NER

We apply Llama and Gemini to the ING data for NER. These results are shown in Table 16. In these tables, the results are divided based on the prompts per model mentioned in Section 4. Because of post-processing difficulties for OpenChat and Llama, we decided to include only Llama due to time constraints, as Llama has shown superior performance compared to OpenChat in sentiment analysis.

Model	Input	English		Multilingual	
		F1 location	F1 organization	F1 location	F1 organization
Llama	Prompt 1: PS	22%	35%	12%	19%
	Prompt 2: Def.	23%	30%	12%	16%
	Prompt 3: Few-shot	11%	11%	7%	12%
Gemini	Prompt 1: PS	**42%**	31%	29%	26%
	Prompt 2: Def.	41%	**40%**	**30%**	**36%**
	Prompt 3: Few-shot	27%	33%	20%	29%

Table 16: Results of applying LLMs for NER on ING news data.

Table 16 shows that the LLMs do not get high performances for NER. With Llama, we see that the more extensive the prompt becomes, the less well the model performs. It seems to be the case that, especially when given some examples, the model's performance declines. On top of that, the output the model gave was not consistent with the output requested in the prompt, nor consistent over the different data points. This makes the output hard to analyze and much less usable.
Gemini shows a better performance than Llama in general. Also, it is notable that for the second, more extensive elaboration prompt for both the English and multilingual setting Gemini performs best. This way, it seems that the extra explanation helps, however, the few shot examples make it more difficult or confusing for the model.
Overall, Gemini performs better. With the NER, we do see for both models that the performance is substantially higher for the English data, compared to the multilingual data. A possible reason for this could be that for location and organization names, these models have seen more English versions in their training data and are thereby more likely to recognize this.

5.3 BERT vs LLMs on financial news

Now, to answer our third and last research question, *How do the performances of BERT models and LLMs differ in sentiment analysis and NER for English and multilingual financial news?*, we compare the performance of the best-performing BERT models and LLMs on the benchmark data and ING financial news data for both sentiment analysis and NER.

5.3.1 Sentiment analysis

Model	F1 negative	F1 neutral	F1 positive	Avg. F1
English test data				
DistilRoBERTa in financial news domain (English)	58.8%	**91.0%**	**87.0%**	78.9%
Gemini prompt 2: definition	**87.9%**	88.6%	76.4%	**84.3%**
Multilingual test data				
DistilBERT multilingual fine-tuned on benchmark	60.5%	65.9%	60.9%	62.4%
Gemini prompt 1: problem statement	**76.4%**	**67.6%**	**67.5%**	**70.5%**

Table 17: Results of best-performing BERT model vs LLMs for sentiment analysis on benchmark data.

In Table 17, we compare the performance of the best-performing BERT models for the test benchmark data to the performance of the best LLM, Gemini, on the benchmark data. In the

first two rows of the table, see the performance of the models on the English benchmark data. In the second two rows, we find the performance of the models on the multilingual benchmark data. This is a repeat of the previously discussed results.
We find that for both the English and the multilingual benchmark data, Gemini outperforms the BERT model, with an overall F1 for English data of 84.3% and 78.9% respectively, and an overall F1 of 70.5% and 62.4% respectively for the multilingual test data. Looking at the English test data, we find tat the BERT model actually outperforms Gemini for neutral and positive, while for negative its performance drops, and that of the LLM is much better. This is likely due to the imbalance of the training data of the BERT model, in which the neutral and positive categories take up the majority. Since the LLM does not have task-specific fine-tuning data, this is not the case for Gemini. Considering the multilingual results, we see a drop in performance for both models. This time, however, the LLM consistently outperforms the BERT model. Overall, we find that the LLM performs best and most consistently. However, not all LLMs outperform all BERT models, but comparing the best LLM and the best BERT model for these datasets, the LLMs do perform better.

Model	F1 negative	F1 neutral	F1 positive	Avg. F1
English test data				
DistilBERT multilingual fine-tuned on benchmark	**87%**	0%	73%	53%
Gemini prompt 2: definition	86%	**46%**	**83%**	**72%**
Multilingual test data				
DistilBERT multilingual fine-tuned on benchmark	69%	7%	61%	46%
Gemini prompt 1: problem statement	**90%**	**54%**	**75%**	**73%**

Table 18: Results of best-performing BERT model vs LLMs for sentiment analysis on ING news data.

In Table 18, we compare the performance of DistilBERT model, fine-tuned on the multilingual benchmark data to the performance of the Gemini model on the ING data. In the first two rows of the table, see the performance of the models on the English ING data. In the second two rows, we find the performance of the models on the multilingual ING data. This is a repeat of the previously discussed results. We find that for both English and multilingual, Gemini substantially outperforms the BERT model, with an overall F1 for English data of 72% and 53% respectively, and an overall F1 of 73% and 45% respectively for the multilingual test data. The only time the BERT model performs better is in the categorization of negative sentiment on the English data. However, this is an increase of only 1%., so not a substantial performance difference. Gemini outperforming BERT on the ING data can be explained by the nature of the data. This data, as explained before when discussing the BERT result, is of a different domain than the data the BERT model had previously seen. For BERT models, it is hard to generalize across domains, while LLMs have a very large cross-domain multilingual training dataset and are hence less prone to circumstances such as language, domain, and task. It is also notable, however, that not all LLMs outperform all BERT models. The size and prompting of the LLM and the newness of the task (in terms of domain, language, or other circumstances) for the BERT model play an important role in their performances.

5.3.2 NER

Model	F1 location	F1 organization	Avg. F1
BERT on news articles domain (English)	**37.0%**	19.0%	**28.0%**
Gemini prompt 2: definition	27.8%	**26.4%**	27.1%

Table 19: Results of best-performing BERT model vs LLMs for named entity recognition on English benchmark data.

In Table 19, we see the performances of the best BERT model for NER on the English benchmark data compared to the performance of the best LLM model and prompt, Gemini using the prompt with an elaborate definition of the problem. The comparison is based only on their performance on the English benchmark data, as we were unable to apply the LLMs to the multilingual benchmark data, as mentioned before. This is a repetition of the previously discussed results. Here, we find that the BERT model outperforms Gemini, with an overall F1 of 28.0% compared to 27.1%, respectively. However, this difference is very small and not substantial, also considering both perform better in a different category. While Gemini is better at finding organizations in the benchmark data, BERT is better at finding locations. In the case of NER on the benchmark data, we cannot conclusively say whether BERT or LLMs perform better.

Model	F1 location	F1 organization	Avg. F1
English test data			
DistilBERT multilingual fine-tuned on benchmark	29%	**55%**	**42%**
Gemini prompt 2: definition	**41%**	40%	41%
Multilingual test data			
DistilBERT multilingual fine-tuned on benchmark	13%	**41%**	27%
Gemini prompt 2: definition	**30%**	36%	**33%**

Table 20: Results of best-performing BERT model vs LLMs for named entity recognition on ING news data.

In Table 20, we see the performances of the best BERT model for NER on the ING data, DistilBERT model, fine-tuned on the multilingual benchmark data, to the performance of the best LLM model and prompt, Gemini using the prompt with an elaborate definition of the problem statement. In the first two rows of the table, see the performance of the models on the English ING data. In the second two rows, we find the performance of the models on the multilingual ING data. This is a repetition of the previously discussed results. For English NER, we find that the BERT model outperforms Gemini, with an overall F1 of 42% compared to 41%, respectively. However, this difference is very small and not substantial, also considering both perform better in a different category. While Gemini is substantially better at finding locations, BERT is substantially better at finding organizations. Looking at the multilingual data, we find that the LLM shows better results than the BERT model with an overall F1 of 33% and 27%, respectively. However, like with the English data, this performance difference is not very substantial and Gemini outperforms the BERT model on locations, while the BERT model outperforms Gemini in finding organizations. In the case of NER, we cannot conclusively say whether BERT or LLMs performs better.

6 Discussion & Limitations

In this study, we conducted an extensive evaluation of the performance of LLMs and BERT on sentiment analysis and NER for multilingual financial news. For this analysis we use the benchmark dataset, and the ING data with 100 news articles. This approach provides insight into the models' generalizability and robustness across different types of text data.
Due to the vast amount of data considered in the merged benchmark data, results of the models on this data are statistically sound. We can trust the result coming from this are not highly influenced by coincidence or outliers. However, due to the merge of data from different domains, we have data with various sentence structures, terminology and other data characteristics. These differences within the data make it more difficult to single out a cause of found patterns in the results.
The ING dataset containing 100 news articles offers insight into the models' real-world applicability for this specific use case. However, due to the small size of the data, it is harder to draw strong conclusions from this. Judging a model's performance on this size dataset is difficult, as for each category there are limited amounts of data points linked to it available, so one wrong prediction can have a big influence on the result and coincidence can play a bigger role as well.
Looking at BERT models versus LLMs, apart from their performance on these datasets for sentiment analysis and NER, there are also some general differences to take into consideration. Firstly, the speed. The speed of both models is dependent on the size of the cluster, GPU power assigned to it and the execution environment. ING works with Google Cloud Platform, which has Gemini integrated in the platform, making the circumstances for running Gemini quite good. Apart from that, all settings were the same for BERT and the LLMs, and yet BERT was, after fine-tuning, over 20 times faster than the fastest LLM (being Gemini). Similarly, there is the energy consumption. Due to the fact that LLMs are of much larger size and complexity, the energy consumption of LLMs for training, as well as per inference, is much higher than that of a BERT model.

6.1 Sentiment analysis

For the sentiment analysis task specifically, there are some limitations related to our data. While the benchmark data for sentiment analysis is usable for robust evaluation due to its size, the disadvantage of this data is that for English, we use financial news, while the multilingual dataset is a mix of domains, causing bias in what the models train and test on. However, an advantage of this characteristic of our data, is that we can test a models ability to transfer knowledge.
With the ING data, there are, apart from the size, also some limitations specifically related to the sentiment analysis task. Firstly, the news articles from ING are a sample taken from their input data that were all consecutive in the data. The data is ordered in time, so these articles are from a relatively short time frame, and not randomly sampled. This is reflected in the data by having some subjects, relevant and thus published in that time frame. This causes a bias and domain specificity that might not properly reflect the whole of ING input data. Moreover, we did labeling for the ING data into negative, neutral, or positive sentiment, knowing that for the result at the moment ING is mostly interested in negative sentiment. We are not experts on the topic, so we had experts check doubtful articles, but not all. Hence there still might be some wrongly labeled articles. Also, after speaking with ING risk managers, we

found that the relevance of an article for them is not solely it being of negative sentiment, as it is defined now. The relevance is more comprehensive and could be covered with sentiment analysis, but would require an extensive labeled dataset for training, testing, and defining what this relevance entails.
Looking at the performance of BERT models on these datasets, we see that the training data is of high influence on the performance of the model on the unseen data. However small the ING dataset is, we do see a substantially different trend in the performance of the models on this data compared to the benchmark data. While on the benchmark data the classification of negative sentiment was the worst, we find that on the ING data the classification of negative sentiment is done best, by different models. This difference indicates that the benchmark data and the ING data are of substantially different structure and domain, causing financially fine-tuned models to stay behind in performance compared to other models fine-tuned on more robust or closer-in-domain data.
Using large language models to perform sentiment analysis works quite well. The models do provide some noise in the output, as, especially the smaller LLMs, seemed unwilling to provide output in the requested manner consistently. However, as the prediction was always provided and can only be one out of three categories, this did not require too much post-processing in order to generate consistently structured results. Looking at the performance, we see that the bigger LLM, in this study being Gemini, works best, and is barely dependent on the prompting. However, it is also notable that smaller models, here Llama, with more extensive prompting can reach competitive performance. Considering this, a larger version of Llama, being openly available, can certainly compete with large commercial models such as Gemini.
When looking at the performance of BERT models compared to the LLMs, we find that for sentiment analysis, LLMs perform better. Many real-world use-cases have data that is not similar to openly available labeled datasets for those tasks. With the ING data for sentiment analysis, this also seems to be the case. The LLMs have a more extensive and diverse train dataset, enabling the LLMs to more easily generalise the sentiment analysis to various domains and contexts. Also, LLMs typically have a larger contextual understanding, more parameters and more layers compared to BERT models, allowing them to be more aware of nuances and other more subtle sentiment cues. All in all, this combined with the limited post-processing required makes LLMs a strong models for sentiment analysis.

6.2 Named entity recognition

Also for NER, there are some task specific considerations. Related to the benchmark data, a difficulty for evaluation matters is the mix in difficulty level, domains and sources of the merged NER datasets. The Wnut17 dataset for example, is known to be very hard as it focuses on identifying unusual, previously unseen entities. CoNLLpp is based on the CoNLL2003 NER dataset, which is a well-known and used lexicon containing news articles on which results of appying NER have been very good. These differences make it harder to evaluate where a model goes wrong, but also tests the robustness of the models.
For the ING data, there are several limitations related to NER to be taken into considerations. The ING data contains quite some noise. The news articles are directly scraped from the web, and hereby not filtered on commercial information in between or at the end of an article. Examples of what we saw in the articles are 'Picture from GettyImages' or 'Share on Twitter / Instagram / Facebook'. These are all company names, that we now required to include in our NER analysis, but are of different appearance than the mention of organizations usually

are in news articles. Also, these entities are unwanted results, as ING wants information on organizations the article is about, not that unrelatedly are mentioned. Moreover, information on publishers causes noise, including sentences such as 'like and subscribe to newsletters from Reuters' or 'published in Australia'. This, again, adds irrelevant organizations and locations to the data. Apart from this, the ING data is not IOB-labeled but has a list of locations and organizations as they occur in the articles instead. However, this list is only available in English. Thereby, the multilingual results provided by multilingual models needs to be translated in order to compare it to the ING labels. This makes it possible that some correctly identified results get lost in translation, as for this we are dependent on the performance of translation models. Lastly, like mentioned for sentiment analysis, the bias in the data on topics discussed in the time frame of the collected articles is also of influence on NER. When diving deeper into the data we find that locations and organizations mentioned also relate to what is in the news right now. For example, relatively many times 'Russia' and 'Ukraine' are locations mentioned, while many countries never appear. The same goes for organizations such as 'Sberbank' and 'Gazprom'. These are some NER biases to take into consideration for the ING data.
Now looking at the performance of BERT models for NER, we see no massive performance shifts comparing the models on the same dataset, nor compared to the ING data. However, it is notable that while organizational classification was lacking on the benchmark data, it's performance caught up and became equal to that of locations classification for the ING data. This indicates that the benchmark data was similarly difficult to the ING data. However, saying the data is similar in other aspects is hard to say as the performance of the models on both datasets is quite low, so it could be finding different structures or having different aspects it is good at for both datasets. Also, not all BERT models have been applied to the benchmark data due to resource limitations. However, all is applied to the ING data. The same goes for LLMs used for NER.
For LLMs, these limitations were extended to the ING data for Llama and OpenChat. Due to inconsistency of the output of those models, they required a lot of time-intensive post-processing, which makes them unusable in a real-world scenario such as for the ING use-case. Hence, we only did this for Llama and do not have results of OpenChat for NER. As for the LLMs performances, we see that the bigger model, being Gemini, outperforms the smaller model, Llama. Also, when given few-shot examples, the performance of the models substantially decreases for both models. Prompt engineering for this task does not seem to improve performance in any extend.
Comparing the performance of BERT and LLMs for NER, we find that there is no substantial difference in performance. NER is dependent on recognizing specific patterns and entities, and extracting those entities exactly from a text. While BERT is proficient at this due to its masked language modeling pre-training and inability to make up text but rather only label each token, LLMs are known to hallucinate and make up a full text, making it more difficult for these models to extract very exact entities from a text. The performance of both models is not high, however, it is notable that these results are the bare minimum as we take every exact entity in a text into consideration. For the ING use-case, finding an entity in a text once would be enough, as well as a 'normalized' version (example: ING instead of ING Group for LLM and only the B-ORG for BERT). This would be beneficial for the actual final performance of both LLMs and BERT models.

7 Conclusion & Future Work

In a nutshell, despite some limitations on our datasets, we find that BERT and LLMs both have advantages and disadvantages that should be taken into consideration when choosing which model to use for a specific use case. To sum up, BERT has difficulties generalizing cross-domain when learning a specific task but can be very strong when fine-tuning and testing data are available in high quantity, quality, and similarity. Also, they are more resource-efficient. LLMs have difficulty with token-level multi-class classification and work better on a text level. However, they need very little to no extra training data and generalize well to unseen domains or use cases. Considering this, we will now answer the proposed research questions.

> **RQ1** How well can BERT models analyze English and multilingual financial news for sentiment analysis and NER?

We find that a BERT model specifically fine-tuned for financial sentiment analysis can very well predict financial sentiment of the same type, as seen with the BERT models fine-tuned and tested on the financial data. However, when adding the multilingual aspect, or changing the domain, those performances rapidly drop. For NER, the domain of the data seems to matter less, as long as the entities are in a similar recognizable structure as the train data. All in all, we conclude that BERT models can be very good at this task when provided train data significantly representative of the use-case, however is strongly data quality dependent.

> **RQ2** How well can LLMs analyze English and multilingual financial news for sentiment analysis and NER?

LLMs have shown to be quite strong in sentiment analysis for multilingual financial news. It shows to strongly benefit from prompt engineering, especially in case of smaller LLMs. NER seems to be a harder task for LLMs comparatively, however, their performance still is consistent over the data and seems to perform better on untokenized input. For NER, the models thus far did not seem to benefit from the same prompt tuning method as applied to sentiment analysis. To conclude, we see that LLMs can be strong at these tasks, though they might need some prompt tuning and post-processing.

> **RQ3** How do the performances of BERT models and LLMs differ in sentiment analysis and NER for English and multilingual financial news?

Lastly, we see that for both models, sentiment analysis seems better doable than the NER task with this data. We conclude that for sentiment analysis, LLMs are stronger with the right prompt, as they are not dependent on training data when applied to domain specific, multilingual or other unforeseen settings. For NER however, the performances of both models are fairly similar, though due to the speed and post-processing required for LLMs, we would say the BERT models are stronger in this sense.
For the ING use-case, we advise to take multiple considerations into account when choosing between BERT and LLM for each task. Firstly, and most obviously, the performance, as extensively discussed before. Secondly, the speed of the models. Considering the high quantity of data coming into the ARIA pipeline, having each step take a long time will lead to inefficient execution. Thirdly, the energy consumption. Apart from speed, each run of an LLM consumes much more energy than a BERT model, which could be an environmental consideration when

performances are similar. Fourth and lastly, the money aspect. There are open-source BERT models and LLMs available which can be good for this use-case. However, not all are accepted by ING policy. For their security policy, a model needs to be approved by the board. So far, Gemini is approved and other models need to be requested per use-case. Gemini is a commercial model from Google, so the price can also be an aspect taken into consideration.
Our advice would be to apply (open-source) LLMs for sentiment analysis with prompt tuning, as these are more robust to unseen cases and topics. If deciding to use BERT, we would advise providing an extensive and robust dataset from the ING data for fine-tuning. For NER, we advise using BERT models, considering the performances are similar, but resource-wise BERT models are more efficient. Also, while the performances of the model seem low in this report, analysis performed with only the entities wanted by ING shows a much more optimistic view. Moreover, the open-source LLMs require extensive post-processing. However, in the case of choosing for LLMs for NER, we would suggest providing the LLM with a list of the entities ING is interested in as a knowledge base. In this way, the LLM can apply NER and entity linking at once looking for specified entities in a text, without having to extract exact entities from a text. An LLM working with a knowledge base, increasing the accuracy of the output, as well as reducing the hallucination problem [LPP+20].
In order to mitigate this work's limitations, there are some suggestions for further enhancement of this study. Firstly, expanding the ING dataset. Using a larger and more diverse datasets will lead to a more comprehensive evaluation of the ING use-case. Secondly, due to the nature of our data, an in-depth error analysis would be beneficial. Conducting detailed error analyses would lead to identification of specific areas where each model excels or fails. Thirdly, resource efficiency studies could be performed. Evaluating the computational costs associated with each model would provide a holistic view of their practicality in different use cases. Lastly, using language detection and having separate models for English and other languages could be a more resource efficient approach worth investigating.

References

[A+13] Piyush Arora et al. Sentiment analysis for hindi language. *MS by Research in Computer Science*, 2013.

[ABB+24] Aref Mahdavi Ardekani, Julie Bertz, Cormac Bryce, Michael Dowling, and Suwan Cheng Long. Finsentgpt: A universal financial sentiment engine? *International Review of Financial Analysis*, 94:103291, 2024.

[AI24] Imtiaz Ahmed and Raisa Islam. Gemini-the most powerful llm: Myth or truth. *Authorea Preprints*, 2024.

[AM21] Shivaji Alaparthi and Manit Mishra. Bert: A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9(2):118–126, 2021.

[Ara19] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[BHV+23] Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[CKJM21] Yuan Chang, Lei Kong, Kejia Jia, and Qinglei Meng. Chinese named entity recognition method based on bert. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, pages 294–299. IEEE, 2021.

[CMST16] Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. Selection criteria for low resource language programs. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[DBdV16] Joel Janek Dabrowski, Conrad Beyers, and Johan Pieter de Villiers. Systemic banking crisis early warning systems using dynamic bayesian networks. *Expert systems with applications*, 62:225–242, 2016.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DLD+22] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[DNvEL17] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[GID+21] Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Abdullah, Rakhi Batra, and Mudasir Ahmad Wani. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490, 2021.

[GMOO10] Dieter Gramlich, Gavin Miller, Mikhail V Oet, and Stephen J Ong. Early warning systems for systemic banking risk: critical review and modeling implications. *Banks and Bank Systems*, 5:199–211, 2010.

[HBR19] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics*, pages 187–196, 2019.

[HJLS13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[HP19] Kai Hakala and Sampo Pyysalo. Biomedical named entity recognition with multilingual bert. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61, 2019.

[hug] Hugging Face – The AI community building the future. — huggingface.co. `https://huggingface.co`. [Accessed 07-07-2024].

[Hus18] Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338, 2018.

[INGa] ING.com. ING at a glance — ing.com. `https://www.ing.com/About-us/ING-at-a-glance.htm`. [Accessed 06-03-2024].

[INGb] ING.com. Wholesale Banking — ing.com. `https://www.ing.com/About-us/Annual-Reporting-Suite/Wholesale-Banking.htm`. [Accessed 06-03-2024].

[KAA+21] Lal Khan, Ammar Amjad, Noman Ashraf, Hsien-Tsung Chang, and Alexander Gelbukh. Urdu sentiment analysis with deep learning methods. *IEEE access*, 9:97803–97812, 2021.

[KAS+21] Asad Khattak, Muhammad Zubair Asghar, Anam Saeed, Ibrahim A Hameed, Syed Asif Hassan, and Shakeel Ahmad. A survey on sentiment analysis in urdu: A resource-poor language. *Egyptian Informatics Journal*, 22(1):53–74, 2021.

[KHX+24] Vipina K Keloth, Yan Hu, Qianqian Xie, Xueqing Peng, Yan Wang, Andrew Zheng, Melih Selek, Kalpana Raja, Chih Hsuan Wei, Qiao Jin, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. *Bioinformatics*, 40(4):btae163, 2024.

[LH20] Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983, 2020.

[LOG+19] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[LPP+20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[LSHL20] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.

[LSP+22] Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi. Mono vs multilingual bert: A case study in hindi and marathi named entity recognition. *arXiv preprint arXiv:2203.12907*, 2022.

[MCH20] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

[ml] meta llama. meta-llama/Meta-Llama-3-8B · Hugging Face — huggingface.co. `https://huggingface.co/meta-llama/Meta-Llama-3-8B`. [Accessed 07-07-2024].

[MSBB21] Loitongbam Sanayai Meetei, Thoudam Doren Singh, Samir Kumar Borgohain, and Sivaji Bandyopadhyay. Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation*, 55(4):947–969, 2021.

[OBGO13] Mikhail V Oet, Timothy Bianco, Dieter Gramlich, and Stephen J Ong. Safe: An early warning system for systemic banking risk. *Journal of Banking & Finance*, 37(11):4510–4533, 2013.

[PM] et al. Pekka Malo. takala/financial_phrasebank · Datasets at Hugging Face — huggingface.co. `https://huggingface.co/datasets/financial_phrasebank`. [Accessed 27-05-2024].

[PNI+18] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, and K Lee. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[Pol] Jean-Baptiste Polle. Jean-Baptiste/financial_news_sentiment · Datasets at Hugging Face — huggingface.co. `https://huggingface.co/datasets/Jean-Baptiste/financial_news_sentiment`.

[PVFE21] Marco Pota, Mirko Ventura, Hamido Fujita, and Massimo Esposito. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Systems with Applications*, 181:115119, 2021.

[Qia] Tay Yong Qiang. GitHub - tyqiangz/multilingual-sentiment-datasets: A collection of multilingual 3-class sentiments (positive, neutral, negative) dataset. — github.com. `https://github.com/tyqiangz/multilingual-sentiment-datasets`.

[R+03] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[RLC19] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*, 2019.

[RNS+18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[RNSS18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[Sau21] Danielle Saunders. *Domain adaptation for neural machine translation*. PhD thesis, 2021.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[SLL+23] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

[SNL19] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[SSR+19] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. Bert for stock market sentiment analysis. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1597–1601, 2019.

[TKSDM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[TMC+21] Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Ver24] Suzan Verberne. Is de zoekmachine van de toekomst een chatbot? https://scholarlypublications.universiteitleiden.nl/handle/1887/3754461, 2024.

[VSP+17] 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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[WCZ+23] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023.

[WMR+20] Zihan Wang, Stephen Mayhew, Dan Roth, et al. Extending multilingual bert to low-resource languages. *arXiv preprint arXiv:2004.13640*, 2020.

[WRK22] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[WSL+19] Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. Crossweigh: Training named entity tagger from imperfect annotations. 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)*, pages 5157–5166, 2019.

[WSL+23] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, 2023.

[WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[WZZZ21] Tongyu Wang, Shangmei Zhao, Guangxiang Zhu, and Haitao Zheng. A machine learning-based early warning system for systemic banking crises. *Applied economics*, 53(26):2974–2992, 2021.

[Xin24] Frank Xing. Designing heterogeneous llm agents for financial sentiment analysis. *arXiv preprint arXiv:2401.05799*, 2024.

[YC14] Bishan Yang and Claire Cardie. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 325–335, 2014.

[ZLZZ21] Lingyun Zhao, Lin Li, Xinhao Zheng, and Jianwei Zhang. A bert based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238, 2021.

[ZOB13] Norshuhani Zamin, Alan Oxley, and Zainab Abu Bakar. Projecting named entity tags from a resource rich language to a resource poor language. *Journal of Information and Communication Technology*, 12:121–146, 2013.

[ZYZ+23] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. *arXiv preprint arXiv:2310.04027*, 2023.

[ZZ23] Yuzhe Zhang and Hong Zhang. Finbert–mrc: Financial named entity recognition using bert under the machine reading comprehension paradigm. *Neural Processing Letters*, pages 1–21, 2023.

[ZZL+23] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.