“Do CoNLL-2003 Named Entity Taggers Still Work Well in 2023?”

Important pts:-

In this paper, we evaluate the generalization of over 20 different models trained on CoNLL-2003, and show that NER models have very different generalization

no evidence of performance degradation in pre-trained Transformers, such as RoBERTa and T5, even when fine-tuned using decades old data.

why some models generalize well to new data while others do not, and attempt to disentangle the effects of temporal drift and overfitting due to test reuse

most deterioration is due to temporal mismatch between the pre-training corpora and the downstream test sets

***Four factors are important for good generalization: model architecture, number of parameters, time period of the pre-training corpus, in addition to the amount of fine-tuning data**.*

performance metrics like accuracy or F1 score on public test sets

These scores are all calculated using the same publicly available test set, which raises several questions. One concern is how much of this progress is actually due to adaptive overfitting, i.e. over-estimating performance by reusing the same test set, as opposed to genuine improvement

there is also the issue of temporal drift as training data ages, which can negatively impact performance on modern data

Performance degradation is a significant concern in applications that use NER, such as text de-identification (Morris et al., 2022), relation extraction (Zhong and Chen, 2021), linking entities to a knowledge base (De Cao et al., 2022), etc.

we created a new test set called CoNLL++. We closely modeled CoNLL++ after the CoNLL-2003 test set, using news articles from 2020 instead of 1996, as in the original dataset.

Using CoNLL++, we conduct an empirical study of more than 20 NER models that were trained on the original CoNLL-2003 training split.

AMBASSADOR	O
TO	O
THE	O
UNITED	I-ORG
NATIONS	I-ORG
:	O
LINDA	I-PER
THOMAS-GREENFIELD	I-PER

Analysis shows that different models can have very different generalization when moving to modern data.

we do not observe evidence of widespread overfitting on CoNLL-2003. On average, each point of F1 improvement on the CoNLL-2003 test set translates to a larger improvement on CoNLL++

Surprisingly, for some models (e.g. RoBERTa and T5), we find no evidence of performance degradation at all, despite the fact they are fine-tuned on a 20-year-old public dataset

The CoNLL-2003 shared task collected English data from the Reuters Corpus, including Reuters news articles published between Aug. 1996 and Aug. 1997. The test set was collected from December 1996

Our dataset follows this distribution to collect Reuters news articles published between December 5th and 7th, 2020, collected from the Common Crawl Foundation

We tokenize the data with the same tokenizer used for the CoNLL-2003 shared task, and randomly select articles to match the total number of tokens in the original test set.

manually labeled this new dataset, which we refer to as CoNLL++, using the BRAT annotation interface

Articles were distributed between two authors, where one author annotated 96.1% of the articles and the other annotated 50.0%. The first author's annotation is used as the gold standard

We find that the CoNLL++ annotations closely follow the style of the original dataset. When considering labels in the CoNLL-2003 test set as gold, our manual reannotation achieves a 95.46 F1 score

The second author's annotation, when considering the first author's as gold, receives a 96.23 F1 score on overlapping articles.

The token-level Cohen's Kappa between the two authors is 97.42, which can be considered almost perfect agreement

We select models with a variety of architectures and pre-training corpora and fine-tune these models to study how different factors affect generalization.

Scripts for fine-tuning Flair and ELMo are adapted from Reiss et al. (2020). 5 Other recurrent neural network (RNN) models are trained using various GitHub

repositories (see footnotes 6, 7 and 8). We fine-tune the BERT and RoBERTa models with the HuggingFace transformers library (Wolf et al., 2020), except LUKE with AllenNLP (Gardner et al., 2018). T5 is fine-tuned to conditionally generate NER tags around entities (e.g. Jane Doe).

{8, 16, 32} and {1e-5, 2e-5, 3e-5, 5e-5} are used for most searches for batch sizes and learning rates respectively

We train models on the CoNLL-2003 training set for 10 epochs, and use the dev set to select the best epoch and other hyperparameters for evaluation. Each model is evaluated five times with different random seeds on the CoNLL-2003 test set and on CoNLL++ to obtain the average F1.

$$\Delta F_1 = \frac{F_1^{\text{CoNLL++}} - F_1^{\text{CoNLL-2003}}}{F_1^{\text{CoNLL-2003}}} \times 100$$

Some models (e.g. RoBERTa-based models and T53B), have no performance drop on CoNLL++, whereas other models' performances decrease significantly

disentangle to what extent the observed performance drops on CoNLL++ are caused by temporal deterioration, or adaptive overfitting.

It has been shown that the size of pre-trained models affects their performance (Kaplan et al., 2020; Raffel et al., 2020).

We observe, from Table 6, that larger models perform better on both test sets, but more impor- tantly, as illustrated in Figure 3, performance degradation on CoNLL++ diminishes or even disappears as the model size grows. The only exception is the RoBERTa-based models, whose base-sized model already achieves comparable performance on CoNLL++.

Figure 3 suggests that larger model sizes not only increase performance on a static test set, but also help models generalize better to new data.

Both RoBERTaLarge and T53B achieve a performance increase of ~0.6%, but the number of parameters of T53B is ~10 times that of RoBERTaLarge. This shows that the generalizability of a model is also affected by factors other than the size of the model, but with the same architectures, larger models tend to generalize better.

The fact that most Transformer-based models achieve higher rankings in CoNLL++ also confirms that pre-trained Transformers generalize better to new data

BiLSTM models with Flair and ELMo embeddings, despite performing exceptionally on CoNLL-2003, show larger performance drops on CoNLL++ (5-6% F1), and the performance of BiLSTM+GloVe models drops even more significantly (greater than 6% F1). Such results show a clear trend that Transformer-based models generalize better to new data.

The generalizability of a model may also be affected by the size of the fine-tuning dataset. We conduct experiments varying the number of CoNLL-2003 training examples used for fine tuning from 10% to 100%.

Both RoBERTaBase and Flair embeddings show improved generalization as we use more training examples. However, this improvement is more pronounced for Flair than RoBERTaBase.

The empirical evidence supports our hypothesis that having more training examples can improve the generalizability of the model, but such effect may vary across different models. RoBERTa-based models generalize well to new data even when only a small amount of fine-tuning data is available, whereas Flair benefits much more from having more fine-tuning data.

Models in Table 2 show different levels of performance drop, or sometimes performance gain, on CoNLL++ compared to the CoNLL-2003 test set, and it is not entirely clear what causes this difference. We hypothesize two potential causes, namely adaptive overfitting (§ 5.1) and temporal drift (§ 5.2).

investigate if the performance drop is caused by adaptive overfitting of models developed over the past 20 years. Roelofs et al. (2019) defined adaptive overfitting as the overfitting caused by reusing the same test set (test reuse)

. Recht et al. (2019) studied this phenomenon in the context of ImageNet by measuring to what extent can the improvement on the old test set translate to improvement on the new test set (diminishing return)

We have not found any diminishing return on CoNLL++, and therefore no adaptive overfitting caused by the model development over the past two decades.

If the models are overfitting to the CoNLL-2003 test set due to test reuse, we should see not only a performance degradation on CoNLL++, but also a performance degradation on a test set taken from the same distribution as the CoNLL-2003 test set.

Most models suffering from performance degradation on the CoNLL++ also perform better on the CoNLL-2003' test set. This provides evidence that individual models are not overfitting to the CoNLL-2003 test set.

Temporal drift refers to the performance degradation of a model on the downstream task caused by the temporal difference between the train and test data.

Prior work has shown that the performance on NER is affected by temporal drift. For example, Rijhwani and Preotiuc-Pietro (2020) showed that the performance of GloVe and Flair embeddings on NER degrades when the test data is more temporally distant from the train data of the downstream task. Agarwal and Nenkova (2022) also reported the same observation on GloVe embeddings

term "temporal drift" but refer to the deterioration of generalization of models caused by the temporal difference between the pre-training corpus of their word embeddings and the test data of the downstream task

We hypothesize that generalization is largely affected by such temporal drift. We conduct experiments on Flair and ELMo, as well as on RoBERTa

having the training corpus for Flair and ELMo embeddings temporally closer to the CoNLL++ test set improves generalization. Notably, the performance gap for ELMo is reduced to -1.43%, better than that of BERTLarge (-2.01%). The improvements in generalization are attributed to the performance drops on the CoNLL-2003 test set and improvements on CoNLL++.

This provides evidence that the generalizability of the LSTM-based contextualized word embeddings is affected by temporal drift. However, even temporally closer data, these models still suffer from performance drops. This suggests that other ingredients, such as model architecture (§ 4.2), are still needed for a good generalization.

temporal drift is the main driving factor for the different levels of generalization.

found that the performance of LMs degrades as the temporal distance between the training data and the test data increases, sometimes called "temporal misalignment" (Luu et al., 2022).

we study performance deterioration on a dataset that has been heavily used to develop NER models over a period of 20 years, and conduct extensive experiments that aim to disentangle the effects of aging training sets from those due to heavy test reuse.

IMP :- *Most closely related to our work is Agarwal and Nenkova (2022), who analyzed a recently created Twitter NER dataset (Rijhwani and Preotiuc-Pietro, 2020) over the period 2014-2019, and found no performance deterioration when using RoBERTabased representations*

Prior work has attempted to mitigate temporal degradation, mostly through continuously updating LMs with new data (Jang et al., 2022; Jin et al., 2022; Loureiro et al., 2022). Luu et al. (2022) explored this idea but found that temporal adaptation is not as effective as fine-tuning on the data from whose time period the dataset is drawn.

In addition, catastrophic forgetting (Robins, 1995) can also be a problem when updating the LMs. Jin et al. (2022) found that applying knowledge distillation (Hinton et al., 2015) based approaches to continual learning can mitigate catastrophic forgetting, while improving the temporal generalization of LMs. Dhingra et al. (2022) proposed to train the LMs with an additional temporal objective by conditioning on the year of data, and found that this effectively mitigated catastrophic forgetting.

we evaluate the generalization of NER models using CoNLL++, a CoNLL-style annotated NER test dataset with data from 2020. We conduct experiments on more than 20 models and find that models exhibit different generalizability. Surprisingly, we find that generalizability is not affected by adaptive overfitting, but rather by temporal drift

To achieve better generalization, we need the combination of four factors: a modern transformer-based architecture, a large number of parameters, a large amount of fine-tuning data and a temporally closer pre-training corpus to the test set.

We find that our progress on developping NER taggers is largely successful, showing not only good performance on individual test set but also good generalization on new data. This allows CoNLL2003 taggers to still work in 2023.

Future research can focus on ways to mitigate temporal drift. Investigation on attributes of pretraining or fine-tuning corpora that causes temporal drift, such as change of entities mentioned, different usage of language, etc., can also shed light

on the more specific impacts from temporal drift, thereby inspiring new and better ways to mitigate it

This paper investigates the generalization of well-named entity recognition (NER) models to data, trained on the popular CoNLL-2003 dataset from 1996. Using data from 2020, the authors develop a new test set called CoNLL++, closely emulating the structure and design of CoNLL-2003. After assessing more than 20 distinct models, they conclude that temporal drift—rather than overfitting to the original dataset—is the main cause of performance deterioration on CoNLL++.Even when trained on decades-old data, transformer-based models like RoBERTa and T5 exhibit outstanding generalization. This implies that these models are more resilient to changes over time.
According to the study, generalization is influenced by a number of parameters, including model size—larger models tend to generalize better—pre-training corpus—more recent data lead to higher performance—and the volume of fine-tuning data—some models' ability to generalize can be enhanced by this amount of data. The paper comes to the conclusion that temporal drift is a major concern even though CoNLL-2003 models continue to perform well in 2023. This emphasizes the necessity of further study to create plans for reducing this impact and guaranteeing the long-term stability of NLP models.

Claims

1. We hypothesize that generalization is largely affected by such temporal drift. We conduct experiments on Flair and ELMo, as well as on RoBERTa

1. Temporal Drift is the Primary Cause of Performance Degradation, not Overfitting:
Support: To carefully control for adaptive overfitting, the authors created a new test set, CoNLL-2003', derived from the same distribution as the original CoNLL-2003 dataset. Their results had most models improving on CoNLL-2003, which indicated that they had not overfitted on the initial test set. Moreover, diminishing

returns analysis—that is, every improvement point obtained on the original CoNLL-2003, translated to an even higher improvement on CoNLL++—did not reveal any sign of overfitting. This is further supported through the large positive correlation observed between the performance improvement on the original dataset and on CoNLL++.

2. Transformer-based Models Exhibit Superior Generalization:
Support: The authors compared the performance of Transformer models, including RoBERTa and T5, to BiLSTM models with different embeddings, such as Flair, ELMo, and GloVe. They demonstrated that for all experiments, Transformer models showed lower performance drops, often less than 4%, on CoNLL++; the same could not be said of the BiLSTM models, where degradation was significant, up to 20% drops. This leads to the possibility that Transformer architectures, through their ability to learn more complicated representations of language, are better at adapting to changes over time.

3. Pre-training Corpus Temporal Proximity is Crucial:
Support: The authors have run experiments where they have fine-tuned models using embeddings pre-trained on different corpora. They have found that training embeddings on a corpus closer in time to the test set greatly improves generalization, reducing the gap. This confirms the effect of temporal drift, as the time distance of the pre-training data directly affects performance in downstream tasks. Moreover, continued pre-training of RoBERTa on data from different years has shown a clear correlation between the date of pre-training data and model generalization. Models that were pre-trained on more recent data generalize better, confirming the importance of temporal proximity.

Authors state that the amount of fine-tuning data affects the generalization ability. Is it the qualitative factors in the data that affects the performance or just the volume of data? (for instance, why does Flair benefit more than RoBERTa in terms of increasing the fine-tuning data)

The paper mentions that Transformers are better at generalizing, mostly due to their ability to learn more complex representations, but the authors don't explain in detail as to why?
What features of Transformer architectures might contribute to their superior generalization capabilities?

1. Agarwal and Nenkova (2022):- analyzed a recently created Twitter NER dataset (Rijhwani and Preotiuc-Pietro, 2020) over the period 2014-2019, and found no performance deterioration when using RoBERTabased representations
2. language modeling (Lazaridou et al., 2021), NER (Augenstein et al., 2017; Agarwal and Nenkova, 2022; Rijhwani and Preotiuc-Pietro, 2020; Ushio et al., 2022), QA (Dhingra et al., 2022), entity linking (Zaporojets et al., 2022), and others (Luu et al., 2022; Amba Hombaiah et al., 2021). All of this work has found that the performance of LMs degrades as the temporal distance between the training data and the test data increases, sometimes called "temporal misalignment" (Luu et al., 2022).

Limitations:-

1. CoNLL++ will not resolve the problem of generalization to modern data. As new data keep emerging, there will always be the question of how well NER models generalize to that new data.
2. Single New Train/Dev/Test Split for Test Reuse Analysis: The analysis of test reuse was conducted on only one new split, potentially limiting the generalizability of the findings. Future studies could repeat this analysis with multiple splits, providing more concrete evidence.
3. Study was focused on only CoNLL-2003. Thus, the conclusions in this paper might not hold true for other NER datasets. Future work would be to extend the hypothesis to other datasets.

(Gemini AI)
(List 2-3 questions you have that you found difficult to understand in the paper.)
Authors state that the amount of fine-tuning data affects the generalization ability. Is it the qualitative factors in the data that affects the performance or just the volume of data? (for instance, why does Flair benefit more than RoBERTa in terms of increasing the fine-tuning data)

The paper mentions that Transformers are better at generalizing, mostly due to their ability to learn more complex representations, but the authors don't explain in detail as to why? What features of Transformer architectures might contribute to their superior generalization capabilities?

“Domain Adaption of Named Entity Recognition to Support Credit Risk Assessment “

Risk assessment is a crucial activity for financial institutions because it helps them to determine the amount of capital they should hold to assure their stability.

Flawed risk assessment models could return erroneous results that trigger a misuse of capital by banks and in the worst case, their collapse

Robust models need large amounts of data to return accurate predictions, the source of which is text-based financial documents.

This paper explores a machine learning approach for information extraction of credit risk attributes from financial documents, modeling the task as a named-entity recognition problem

We propose a solution for domain adaption for NER based on out-of-domain data, coupled with a small amount of in-domain data.

We also developed a financial NER dataset from publicly-available financial documents.

one of the main causes of the GFC was the use of poor financial models in risk assessment

Financial documents such as contracts and loan agreements provide the information required to perform the risk assessment. . These texts hold relevant details that feed into the assessment process, including: ***the purpose of the agreement, amount of loan, and value of collateral.***

bank staff manually extract the information from such financial documents, but the task is expensive and time-consuming for three main reasons: (1) all documents are in unstructured, textual form; (2) the volume of "live" documents is large, numbering in the millions of documents for a large bank; and (3) banks are continuously adding new information to the risk models, meaning that they potentially need to extract new fields from old documents they have previously analyzed

The primary focus of this paper is how to build supervised NER models to extract information from financial agreements based on pre-existing out-of-domain data with partially-matching labelled data, and small amounts of in-domain data.

There are few public datasets in the financial domain, due to the privacy and commercial value of the data.

This paper describes an approach for domain adaption that includes a small amount of target domain data into the source domain data

Related works:-

Farmakiotou et al. (2000) extract entities from financial news using grammar rules and gazetteers. This rule-based approach obtained 95% accuracy overall, at a precision and recall of 78.75%. Neither the number of documents in the corpus nor the number of annotated samples used in the work is mentioned, but the number of words in the corpus is 30,000 words for training and 140,000 for testing.The approach involved the creation of rules by hand; this is a time-consuming task, and the overall recall is low compared to other extraction methods.

The approach of Farmakiotou et al. (2000) is similar to our approach in that they tried to address an NER problem with financial data. However, their data came from financial news rather than the financial agreements, as targeted in our work.

> a rule-based approach was proposed by Sheikh and Conlon (2012) for extracting information from financial data (combined quarterly reports from companies and financial news) with the aim of assisting in investment decision-making. The rules were based on features including exact word match, part-of-speech tags, orthographic features, and domain-specific features. After creating a set of rules from annotated examples, they tried to generalize the rules using a greedy search algorithm and also the Tabu Search algorithm. They obtained the best performance of 91.1% precision and 83.6% recall using the Tabu Search algorithm.

The focus of Sheikh and Conlon (2012) is closer to that in this paper, in that they make use of both financial news and corporate quarterly reports. However, their extraction task does not consider financial contracts, which is the key characteristic of our problem setting.

> Moens et al. (1999) used information extraction to obtain relevant details from Belgian criminal records with the aim of generating abstracts from them. The approach takes advantage of discourse analysis to find the structure of the text and linguistic forms, and then creates text grammars. Finally, the approach uses a parser to process the document content. Although the authors do not present results, they argue that when applied to a test set of 1,000 criminal cases, they were able to identify the required information.
>
> Domain adaptation for named entity recognition techniques has been explored widely in recent years. For instance, Jiang and Zhai (2006) approached the problem by generalizing features across the source and target domain to way avoid overfitting. Mohit and Hwa (2005) proposed a semi-supervised method combining a naive Bayes classifier with the EM algorithm, applied to features extracted from a parser, and showed that the method is robust over novel data.

Blitzer et al. (2006) induced a correspondence between features from a source and target domain based on structural correspondence learning over unlabelled target domain data.
Qu et al. (2015) showed that a graph transformer NER model trained over word embeddings is more robust cross-domain than a model based on simple lexical features.
using a modest amount of labelled in-domain data to perform transfer learning can substantially improve classifier accuracy (Duong et al., 2014).

Named entity recognition (NER) is the task of identifying and classifying token-level instances of named entities (NEs), in the form of proper names and acronyms of persons, places or organizations, as well as dates and numeric expressions in text (Cunningham, 2005; Abramowicz and Piskorski, 2003; Sarawagi, 2008). In the financial domain, example NE types are LENDER, BORROWER, AMOUNT, and DATE.

*We build our supervised NER models using **conditional random fields (CRFs)**, a popular approach to sequence classification (Lafferty et al., 2001; Blunsom, 2007).*

CRFs model the conditional probability p(s|o) of labels (states) s given the observations o as in Equation 1, where t is the index of words in observation sequence o, each k is a feature, wk is the weight associated with the feature k, and Zw(o) is a normalization constant.

$$p(s|o) = \frac{\exp(\sum_t \sum_k w_k f_k(s_{t-1}, s_t, o, t))}{Z_w(o)} \quad (1)$$

We annotated a dataset of financial agreements made public through U.S. Security and Exchange Commission (SEC) filings. Eight documents (totalling 54,256 words) were randomly selected for manual annotation, based on the four NE types provided in the CoNLL-2003 dataset: LOCATION (LOC), ORGANISATION (ORG), PERSON (PER), and MISCELLANEOUS (MISC). The annotation was carried out using the Brat annotation tool (Stenetorp et al., 2012). All documents

were pre-tokenized and part-of-speech (POS) tagged using NLTK (Bird et al., 2009). As part of the annotation, we automatically tagged all instances of the tokens lender and borrower as being of entity type PER.

For the training set, we use the CoNLL-2003 English data, which is based on Reuters newswire data and includes part-of-speech and chunk tags (Tjong Kim Sang and De Meulder, 2003).

The eight financial agreements were partitioned into two subsets of five and three documents, which we name "FIN5" and "FIN3", respectively. The former is used as training data, while the latter is used exclusively for testing.

- Word features: the word itself; whether the word starts with an upper case letter; whether the word has any upper case letters other than the first letter; whether the word contains digits or punctuation symbols; whether the word has hyphens; whether the word is all lower or upper case.
- Word shape features: a transformation of the word, changing upper case letters to *X*, lower case letters to *x*, digits to *0* and symbols to *#*.
- Penn part-of-speech (POS) tag.
- Stem and lemma.
- Suffixes and Prefixes of length 1 and 2.

We first trained and tested directly on the CoNLL2003 data, resulting in a model with a precision of 0.833, recall of 0.824 and F1-score of 0.829 (Experiment1), competitive with the start-of-the-art for the task.

The next step was to experiment with the financial data. For that, first we applied the CoNLL2003 model directly to FIN3. Then, in order to improve the results for the domain adaption, we trained a new model using the CONLL +FIN5 data set, and test this model against the FIN3 dataset

With out-of-domain test data, a precision of 0.247 and a recall of 0.132 (Experiment2) was observed, while testing with in-domain data achieved a precision of 0.833 and recall of 0.824 (Experiment 1)

the source domain (CONLL) data a small amount of the target domain data — i.e. including data from FIN5— generating a new training data set (CONLL +FIN5). When trained over this combined data set, the results increased substantially, obtaining a precision of 0.828, recall of 0.770 and F-score of 0.798 (Experim

Name	Description
CoNLL	CoNLL-2003 training data
$\text{CoNLL}_{\text{test}}$	CoNLL-2003 test data
CoNLL +Fin5	CoNLL-2003 training data + five financial agreements
Fin5	Five financial agreements
Fin3	Three financial agreements

Table 1: Description of the data sets used.

Name	Training Data	Test Data	P	R	F1
Experiment1	CoNLL	$\text{CoNLL}_{\text{test}}$	0.833	0.824	0.829
Experiment2	CoNLL	Fin3	0.247	0.132	0.172
Experiment3	CoNLL +Fin5	Fin3	0.828	0.770	0.798
Experiment4	Fin5	Fin3	0.944	0.736	0.827

Table 2: Results of testing over the financial data sets.

We can see that the more financial data we add, the more the F-score improves, with a remarkably constant absolute difference in F-score between the two experiments for the same amount of in-domain

Analysis of the errors in the confusion matrix reveals that the entity type MISC has perfect recall over the financial dataset. Following MISC, PER is the entity type with the next best recall, at over 0.9. However, generally the model tends to suffer from a high rate of false positives for the entities LOC and ORG, affecting the precision of those classes and the overall performance of the

One interesting example of error in the output of the model is when the tokens refer to an address. One example is the case of 40 Williams Street, where the correct label is LOC but the model predicts the first token (40) to be NANE and the other two tokens to be an instance of PER (i.e. Williams Street is predicted to be a person)

In the model generated with just the CONLL data, one notable pattern is consistent false positives on tokens with initial capital letters; for example, the model predicts both Credit Extensions and Repayment Period to be instances of ORG, though in the gold standard they don't belong to any entity type. ***This error was reduced drastically through the addition of the in-domain financial data in training, improving the overall performance of the model.***

NaNE:- not a named entity

Our experimental results showed that, for this task and our proposed approach, small amounts of indomain training data are superior to large amounts of out-of-domain training data, and furthermore that supplementing the in-domain training data with out-of-domain data is actually detrimental to overall performance.

In future work, we intend to test this approach using different datasets with an expanded set of entity types specific to credit risk assessment, such as values and dates. An additional step would be carry out extrinsic evaluation of the output of the model in an actual credit risk assessment scenario. As part of this, we could attempt to identify additional features for risk assessment, beyond what is required by the financial authorities.

Q1

This paper addresses the challenge of extracting critical credit risk information from financial documents such as loan agreements. The authors use Named Entity Recognition to make the process automatic, as it is manually time consuming and expensive, but indicates a necessity for large amounts of labeled data for training. So, to overcome this, they address domain adaptation—mixing out-of-domain data,

(CoNLL-2003 English data) with a much smaller amount of manually annotated in-domain financial agreements. Their experiments demonstrate that while adding a small amount of in-domain data can improve performance, training on in-domain data alone leads to the best results. This underlines the crucial role of domain-specific information for successful NER in financial contexts.
The contributions to the study include the making of a publicly available dataset of annotated financial agreements for future research. It also stresses the importance of even a small amount of in-domain data for NER tasks and suggests future directions like expanding the entity types (values, dates), conducting real-world evaluations, and identifying additional features relevant to credit risk assessment.

Q2

Out-of-domain data alone is insufficient for achieving good performance in NER tasks specific to the financial domain.

support:- The authors directly apply a NER model trained on the CoNLL-2003 dataset (out-of-domain) to their financial agreement dataset. The precision, recall and F1 scores were very low, compared to models trained with in-domain data, which shows the limitations of relying only on out-of-domain data.

Incorporating a small amount of in-domain data significantly improves the performance of NER models in financial domains.

Support:- The authors experiment with a mixed training set containing both CoNLL-2003 data and a small number of manually annotated financial agreements (Fin5). This approach shows a substantial improvement in performance compared to using only out-of-domain data, which shows the value of even a limited amount of in-domain data.

Training exclusively on in-domain data (financial agreements) yields the best performance for NER tasks in the financial domain.

Support:- The authors compare the performance of their mixed training set model with a model trained solely on in-domain data(only on Fin5 dataset). The purely in-domain model outperforms the other 3 combinations of CoNLL and Fin5 datasets, indicating that the domain-specific content is important for optimal performance in financial NER tasks.

Q3

1. The feature engineering performed on the financial documents to get the "Fin5", "Fin3" datasets wasn't clearly mentioned. How does that affect the results differently?
2. Are there any sub-categories in 'MISC' named entities? If so, how are they classified into different types based on the relevance w.r.t financial data?

Q4

1. Farmakiotou et al. (2000) - The approach of Farmakiotou et al. (2000) is similar to the approach in the given paper. They tried to address an NER problem with financial data. However, their data came from financial news rather than the financial agreements, as targeted in the given paper.
2. Sheikh and Conlon (2012) - They make use of both financial news and corporate quarterly reports. However, their extraction task does not consider financial contracts, which is the main focus of the given paper.

   The paper extends these works by introducing a supervised-ML approach of CRF (conditional random fields), focusing on financial agreements and exploring domain adaptation techniques.

Q5

1. The dataset size was limited, as pointed out by the authors due to the privacy and commercial value of the data.
   Future work could involve expanding the dataset by annotating a larger corpus of financial documents, potentially including different types of agreements and financial reports.
2. The given paper focuses only on a handful of named entities like LOC, ORG, PER, MISC.
   Future work could involve more entity types related to credit risk assessment.

FiNER-ORD: Financial Named Entity Recognition Open Research Dataset

Most datasets created for named entity recognition (NER) are not domain-specific. The finance domain presents specific challenges to the NER task and a domain-specific dataset would help push the boundaries of finance research.

In our work, we develop the first high-quality English Financial NER Open Research Dataset (FiNERORD). We benchmark multiple pre-trained language models (PLMs) and large-language models (LLMs) on FiNER-ORD

The FiNER-ORD dataset will open future opportunities to use FiNER-ORD as a primary benchmark for financial domain-specific NER and NLP tasks.

The abundance of financial texts, primarily in the form of news and regulatory writings, presents a valuable resource for analysts, researchers, and even individual investors to extract relevant information

processes for information retrieval must be efficient to make data driven decisions in a timely manner.

Named entity recognition (NER) is one such NLP technique that serves as an important first step to identify named entities, such as persons, organizations, and locations, and efficiently use available text data to ultimately drive downstream tasks and decisions.

While numerous studies have constructed annotated datasets (Sang and De Meulder, 2003; Derczynski et al., 2017a; Weischedel et al., 2013) and developed NER models (Li et al., 2019; Yamada et al., 2020; Wang et al., 2021a,b) for generic texts, the financial domain presents unique challenges that require domain-specific approaches and expertise

The current annotated dataset based on Credit Risk Agreements (Alvarado et al., 2015) for financial NER has significant shortcomings, limiting the use of deep learning models for financial NER

Our work adds value to the field by establishing the largest Financial Named Entity Recognition Open Research Dataset (FiNER-ORD). Our objective in this context is not to introduce a novel state-of-the-art (SOTA) model architecture. Instead, we undertake a benchmarking exercise, evaluating pre-trained language models (PLMs) and zero-shot large-language models (LLMs) using the FiNER-ORD dataset.

FiNER-139 from Loukas et al. (2022) is an entity recognition dataset for numerical financial data. they also don't use numerical entities because they can be easily extracted using regular expressions and are the easiest entity types to recognize

work in recognizing entities in invoices, business forms, and emails (Francis et al., 2019). CRA NER dataset from Alvarado et al. (2015) was the first attempt to create a financial NER dataset similar to FiNER-ORD, but in our assessment presented in Section 3.2, it is skewed and thus limited in usefulness.

Sentiment analysis of news aids event ranking based on market factors, aiding price prediction (Feng et al., 2021). Analyzing tweets and headlines automates trading, price movement, and risk forecasts (Sawhney et al., 2021).

Extracting claims from analysts' reports improves volatility forecasting on release and earning dates (Shah et al., 2024). Real-time web tools extract operating segments, aiding performance analysis (Ma et al., 2020). Transformer models fine-tuned for price-change data extraction measure inflation exposure (Chava et al., 2022)

Financial knowledge graphs from news data enhance algorithmic trading (Cheng et al., 2020).

FiNER-ORD consists of a manually annotated dataset of financial news articles (in English) collected from webz.io1 . In total, there are 47,851 news articles available for download in this dataset at the time of writing this paper.

For the manual annotation of named entities in financial news, we randomly sampled 220 documents from the entire set of news articles. We observed that some articles were empty in our sample, so after filtering the empty documents, we were left with a total of 201 articles.

Each token is labeled as one of 4 broad entity types: PER, LOC, ORG, and O.

each of the PER, LOC, and ORG classes are further segmented with the suffixes _B (denoting beginning token of a span) and _I (denoting intermediate token of a multi-token span)

Unfortunately, the annotation methodology for the CRA NER dataset "automatically tagged all instances of the tokens lender and borrower as being of entity type PER" (Alvarado et al., 2015). This approach is problematic because of the resulting skewed distribution of entity types in the dataset, leading to confounded results

Our analysis of the CRA NER dataset showed that in FIN3 (CRA test data split), instances of the tokens lender and borrower represented 83.05% of all PER tokens and 44.95% of all tokens labeled as PER, ORG, MISC, or LOC.

Thus, we believe the CRA dataset is not a high quality benchmark for specialized NLP tasks in the financial domain, motivating us to create a new high-quality financial domain-specific NER dataset

The CoNLL NER dataset (Sang and De Meulder, 2003) was created with manual annotation on Reuters generic news stories which were pre-tokenized and part-of-speech (POS) tagged by the memory-based MBT tagger (Daelemans et al., 2002).

financial texts differ from general texts and contain a higher ratio of organization tokens and entities compared to person and location tokens and entities

the finance-specific usefulness of FiNER-ORD over CoNLL is highlighted by the average length values for ORG entities. We see similar average lengths for ORG entities in FiNER-ORD and CRA, both of which are financial NER datasets, but the average length of ORG entities in CoNLL is much smaller

in financial texts, organization entities are more likely to span over multiple tokens. Having a higher percentage of useful ORG entities and tokens can also enhance applications like measuring correlations in financial news networks and market movements (Wan et al., 2021).

We benchmark FiNER-ORD with several base and large transformer-based models. For the base model category, we use BERT (Devlin et al., 2018), FinBERT (Yang et al., 2020), and RoBERTa (Liu et al., 2019). For the large model category, we use BERT-large (Devlin et al., 2018) and RoBERTalarge (Liu et al., 2019). We do not pre-train these models before fine-tuning them

To benchmark the performance of current SOTA generative LLMs, we measure the zero-shot performance on three train validation splits for the "gpt-4- 0613" and "gpt-3.5-turbo-0613" models3 with 0.0 temperature and 1000 max tokens for the output.

Although zero-shot GPT-4 outperforms zeroshot ChatGPT-3.5-Turbo overall, fine-tuned PLMs outperform both zero-shot LLMs across all entity label categories

This finding aligns with the survey by Pikuliak (2023), which finds that zero-shot ChatGPT fails to outperform fine-tuned models on more than 77% of NLP tasks. RoBERTa-base achieves best weighted F1 score overall and for all entity label categories when compared to all models

A future ablation study of these language models would provide key insights such as why RoBERTa-base, a model more generally trained with the masked language modeling (MLM) objective, outperforms (when considering our financial NER evaluation) FinBERT-base, a model trained on financial sentiment classification tasks and widely used for various financial domain-specific NLP tasks

This study evaluates the transfer learning capabilities of the best-performing model, RoBERTa-large, originally trained and tested on the FiNER-ORD dataset.

Specifically, we investigate its performance when fine-tuned on the CRA NER dataset and CoNLL dataset, followed by testing on the test split of the FiNER-ORD dataset.
These results underscore the significance of the FiNER-ORD dataset in enhancing model performance relative to the existing CRA and CoNLL NER datasets

We demonstrate the importance of our FiNERORD dataset compared to the existing CoNLL2003 and financial CRA NER datasets

The performance analysis shows that RoBERTa-base outperforms all tested models (including FinBERTbase) overall, and both GPT-4 and Chat-GPT-3.5- Turbo underperform the tested PLMs over all entity label categories.

We acknowledge a geographic bias in the dataset as FiNER-ORD only includes English financial news articles. In the future, other forms and languages of financial texts could be manually annotated to expand FiNER-ORD. Additionally, more label classes such as 'product', 'miscellaneous', and forms of relationships could be annotated for future downstream applications such as knowledge graph creation.

primary focus of this work is to present the first high-quality financial NER open research dataset and benchmark our FiNER-ORD with multiple PLMs and LLMs to evaluate the performance of these models on the finance domain-specific NER task

We found that SEC Form 10-K filings, filed annually by public companies in the United States, have certain limitations. The 10-K provides details about a company's business, risks, and operating and financial results for each fiscal year. For the NER task, however, we observe there is a noticeable bias towards organization entity tokens associated with the company filing the 10-K. As a result, there is a reduced diversity of named entities, particularly concerning other organizations. Furthermore, the average length of 10-K filings has been growing significantly every year with at least 60,000 words between 2022 and 2023

In contrast, financial news articles are shorter and can be used to better capture the interdependence of companies and the complexity of modern financial markets.

Both annotators had a manual annotation agreement of approximately 96.85% of the 5546 entities across the PER, LOC, ORG entity classes and train, validation, test splits of FiNER-ORD.

The annotation guide was developed iteratively during the annotation process. Online resources were consulted if the annotation guide did not address a specific disagreement, and the annotation guide was updated accordingly afterwards.

To correct potential errors in manual annotations, we run a custom post-processing script that performs the following four tasks: (1) remove trailing spaces from annotated entities, (2) extend token level borders to non-space characters to change an erroneous span to the correct span, (3) clean entity suffixes with techniques such as removing an apostrophe followed by the letter s ('s) from entity suffixes, (4) tokenize text with Stanza (Qi et al., 2020) and add positional information for labeled entities by splitting multi-token spans into separate tokens, assigning _B as the label suffix for the first separated token in the

multi-token span, and assigning _I as the label suffix for the remaining separated tokens in the multi-token span

all tokens which are not annotated with one of PER_B, PER_I, LOC_B, LOC_I, ORG_B, ORG_I are assigned the label O, denoting "other" type of token not belonging to the person, location, organization classes.

Each news article is available in the form of a JSON document with various metadata information including the source of the article, publication date, author of the article, and the title of the article. Entities of the type person (PER), organization (ORG), and location (LOC) were identified according to the rules described

PER entities were identified by their first name and/or last name

bold spans represent a single person entity. In the case where a person was identified by their first and last name, the entire name was labeled as PER and the post-processing script tagged the first name as PER_B and the last name as PER_I

Since the dataset is comprised of financial news articles from July to October 2015, the dataset has a bias towards the news of that time period. An advantage of FiNER-ORD is that it has a unique heterogeneity due to it being composed of English language financial articles published by institutions from around the world unlike the CRA dataset which are from United States Securities and Exchange Commission documents. The most common LOC and ORG entities reflect the global news article source.

There were a few edge cases in the annotation process. For example, labeling entities when location (LOC) is part of the organization (ORG) entity is a common problem in finance. The phrase "Google India" has "India" which is a location, but it is labeled as an organization in our framework. This is because our process does not permit overlapping entity labels. Another such example is "New York Stock Exchange" which we annotated entirely as an ORG entity, despite models often

predicting “New York” as LOC. Therefore, the correct labels are ORG_B for “New” and ORG_I for “York,” whereas models might label them as LOC_B and LOC_I, respectively. Commonly used words like "the" which are often tagged as O (Other) may sometimes be present in the name of an organization, such as "The Wall Street Journal". In such cases, we have tagged "The" to be included as part of an ORG entity.

We use a maximum of 100 epochs for training with early stopping criteria.

In order to test whether FiNER-ORD and CoNLL can be used together to achieve better NER performance, we combine the training data from both datasets. We test separately on the FiNER-ORD and CoNLL test sets. The Weighted Avg. results shown in Table 6 suggest that when testing on CoNLL, FiNER-ORD can be used to complement CoNLL.

Q1

This paper introduces FiNER-ORD, a new, manually annotated dataset for Financial Named Entity Recognition. The dataset is designed to improve the drawbacks in existing financial and general NER datasets with a large corpus of English financial news articles annotated independently with high accuracy by two annotators. Unlike previous datasets, which have very skewed distributions of entities and are small, FiNER-ORD captures several aspects of financial language in a well-represented and varied sample. The authors benchmark several pre-trained language models and large language models on this dataset, finding that RoBERTa-base outperforms other models, including FinBERT, specifically designed for finance, and zero-shot LLMs like GPT-4. Further transfer learning studies underline the representation of financial language quality in FiNER-ORD compared to general datasets.
It highlights the critical nature of domain-specific datasets such as FiNER-ORD in the enhancement of NLP models applied in finance. It acknowledges the limitations of the dataset, focusing on English news articles, and suggests further work on scope extension. The authors conclude that FiNER-ORD offers a valuable

resource for the advance of research and development of financial NLP, paving the path for models that are more robust and accurate to the complexities of the financial domain.

Q2

zero-shot LLMs can perform reasonably well on financial NER but fine-tuned PLMs outperform them.
Support: The authors test GPT-4 and ChatGPT-3.5-Turbo in a zero-shot setting, where they are prompted to identify entities without specific training. While these LLMs achieve some performance, their results are lower than those of fine-tuned PLMs on all entity types. This suggests that while LLMs are powerful, fine-tuning models for financial NER remains crucial for achieving optimal performance.

Existing financial NER datasets perform poorly, thus motivating the need for a new high-quality dataset like FiNER-ORD.
Support: The annotation methodology for the CRA NER dataset "automatically tagged all instances of the tokens lender and borrower as being of entity type PER" (Alvarado et al., 2015). This approach is problematic because of the resulting skewed distribution of entity types in the dataset, leading to confounded results. The authors also argue that general NER datasets like CoNLL-2003 lack the specific focus on financial language and entities, making them less effective for financial domain tasks. The authors also point out that by the transfer learning ablation study, the models trained on CRA or CoNLL perform poorly on FiNER-ORD's test set, highlighting the need for a more relevant dataset.

Q3

What is the rationale for choosing RoBERTa-base as the best-performing model, even though it wasn't specifically trained on financial data?
The paper states that RoBERTa-base outperforms FinBERT, which was specifically trained on financial sentiment classification tasks. A more in-depth analysis of why RoBERTa-base performs well, even without financial domain training, is instrumental in this domain.

How was the transfer learning study conducted using CRA and CoNLL datasets?
The paper states that RoBERTa-large was fine-tuned on these datasets and then tested on FiNER-ORD. However, it doesn't clearly explain how the models were adapted to the different dataset formats or entity labels. A more detailed explanation of the transfer learning methodology is required.

Q4

While numerous studies have constructed annotated datasets (Sang and De Meulder, 2003; Derczynski et al., 2017a; Weischedel et al., 2013) and developed NER models (Li et al., 2019; Yamada et al., 2020; Wang et al., 2021a,b) for generic texts, the financial domain presents unique challenges that require domain-specific approaches and expertise

"Domain Adaptation of Named Entity Recognition to Support Credit Risk Assessment" by Alvarado et al. (2015): This paper introduces the CRA NER dataset, which was the first attempt at a domain-specific dataset for financial NER. However, it suffers from skewed entity distributions and limited size.

This paper overcomes the limitations of the CRA dataset by creating a larger, more diverse dataset of financial news articles with manual annotation. It expands on Loukas et al.'s work by focusing on non-numeric financial entities (like persons, locations, organizations and others), which are crucial for deeper financial understanding.

Q5

The given dataset is limited to only 4 entity types:- PER, ORG, LOC and O
more label classes such as ‘product’, ‘miscellaneous’, and forms of relationships could be annotated for future downstream applications such as knowledge graph creation. Expanding the entity types to include things like products, financial instruments, or events would enhance its usability for more complex financial analysis tasks.

Language Bias: FiNER-ORD is limited to English language financial news articles. This limits its applicability to other languages and potentially introduces biases related to global economic contexts. Expanding the dataset to include multiple languages, particularly those relevant to global finance, would enhance its value and broaden its applicability. This could involve collaborating with researchers and annotators from different geographical regions.