# WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain

**Raj Sanjay Shah**[†] , **Kunal Chawla**[†*] , **Dheeraj Eidnani**[†*] , **Agam Shah**[†*] , **Wendi Du**[†]
**Sudheer Chava**[†] , **Natraj Raman**[♣] , **Charese Smiley**[♣] , **Jiaao Chen**[†] , **Diyi Yang**[♡]
[†] Georgia Institute of Technology
[♣] JPMorgan AI Research
[♡] Stanford University

## Abstract

Pre-trained language models have shown impressive performance on a variety of tasks and domains. Previous research on financial language models usually employs a generic training scheme to train standard model architectures, without completely leveraging the richness of the financial data. We propose a novel domain specific Financial LANGuage model (FLANG) which uses financial keywords and phrases for better masking, together with span boundary objective and in-filing objective. Additionally, the evaluation benchmarks in the field have been limited. To this end, we contribute the Financial Language Understanding Evaluation (FLUE), an open-source comprehensive suite of benchmarks for the financial domain. These include new benchmarks across 5 NLP tasks in financial domain as well as common benchmarks used in the previous research. Experiments on these benchmarks suggest that our model outperforms those in prior literature on a variety of NLP tasks. Our models, code and benchmark data are publicly available on Github and Huggingface[1]

## 1 Introduction

Efficient financial markets incorporate all price relevant information available to investors at that point of time. Unstructured data, such as textual data, help complement structured data traditionally used by investors. For example, in addition to quantitative data such as firm's financial performance, the tone and sentiment of firms' financial reports, earnings calls and social media posts can also influence the stock price movement (Bochkay et al., 2020). We aim to capture these textual features with the help of pre-trained deep learning models, which have shown superior performance in a variety of Natural Language Processing (NLP) tasks (Radford et al., 2019; Devlin et al., 2018; Liu et al., 2019; Lewis et al., 2020). However, the language used in finance and economics is likely to be different from the language of common usage. A statement like "*The crude oil prices are going up*" has a negative sentiment for the financial markets, but it does not contain traditionally negative words such as danger, hate, fear, etc. (Loughran and McDonald, 2011). Therefore, it is necessary to develop a domain-specific language model training methodology that improves the performance in the downstream NLP tasks like managers' sentiment analysis and financial news classification.

Previous research, for example, Yang et al. (2020); Araci (2019) have pre-trained the state-of-the-art language models like BERT (Devlin et al., 2018) with financial documents, but suffer from two major limitations. First, financial domain knowledge and adaptation are not utilized in the pre-training process. We argue that the *financial terminologies* play a critical role in understanding the language used in financial markets, and expect a performance improvement after incorporating the financial domain knowledge into the pre-training process. Second, the lack of different evaluation benchmarks limit the test the language models' performance in finance-related tasks.

In this work, we propose a simple yet effective language model pre-training methodology with preferential token masking and prediction of phrases. This helps capture the fact that many financial terms are actually multi-token phrases, such as *margin call* and *break-even analysis*. We contribute and make public two language models trained using this technique. Financial LANGuage Model (`FLANG-BERT`) is based on BERT-base architecture (Devlin et al., 2018), which has a relatively

Email IDs of the authors: {rajsanjayshah, kunalchawla, deidnani, ashah482, wendi.du, schava6, jchen896}@gatech.edu, natraj.raman@jpmorgan.com, charese.h.smiley@jpmchase.com, diyiy@cs.stanford.edu
* These authors contributed equally to this work
[1]The website can be found at https://salt-nlp.github.io/FLANG/. All the FLANG models are available on the Huggingface SALT-NLP site.

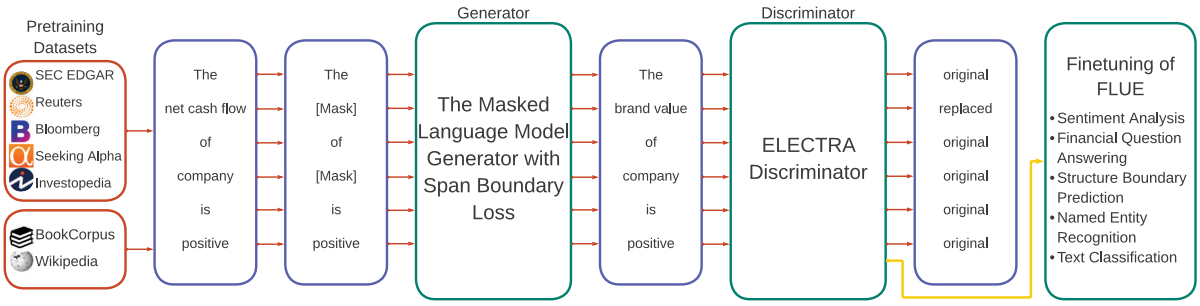


Figure 1: Architecture of our model. We use finance specific datasets and general English datasets (Wikpedia and BooksCorpus) for training the model. We follow the training strategy of ELECTRA (Clark et al., 2020) with span boundary task which first predicts masked tokens using language model and then uses a discriminator to assess if a token is original or replaced. The generator and discriminator are trained end-to-end, and both words and phrases from financial vocabulary are used for masking. The final discriminator is then fine-tuned on individual tasks on our contributed benchmark suite, Financial Language Understanding Evaluation (FLUE). Note that our method is not specific to ELECTRA and can be generalized to other models.

small memory footprint and inference time. It also enables comparison with previous works, most of which are based on BERT. We also contribute `FLANG-ELECTRA`, our best performing model, based on the ELECTRA-base architecture (Clark et al., 2020), where we introduce a span boundary objective on the ELECTRA generator pre-training task to learn robust financial multi-word representations while masking contiguous spans of text. We show that FLANG-BERT outperforms all previous works in nearly all our benchmarks, and FLANG-ELECTRA further improves the performance giving two new state-of-the-art models. Our training methodology can be extended to other domains that would benefit from domain adaptation.

Financial domain benchmarks are critical to evaluate the newly developed financial language models. Inspired by GLUE (Wang et al., 2018), a set of comprehensive benchmarks across multiple NLP tasks, we construct Financial Language Understanding Evaluation (`FLUE`) benchmarks. FLUE consists of 5 financial domain tasks: financial sentiment analysis, news headline classification, name entity recognition, structure boundary detection, and question answering. We intend for this benchmark suite to be a standard for evaluation of natural language tasks in financial domain, subject to appropriate license and privacy considerations. All proposed benchmarks will be made publicly available on Github and Huggingface.

Our contributions are as follows:

- We propose masking finance-specific words and phrase masking for pre-training language model, as well as a span boundary objective to build robust multi-word representations.
- We contribute finance-related benchmarks with 5 NLP tasks: financial sentiment analysis, news headline classification, named entity recognition, structure boundary detection, question answering. This results in a comprehensive suite of finance benchmarks, with licensing details in Table 1.
- We make all our models and code publicly available, for easier development and further research by the NLP and Finance community. Specifically, we contribute FLANG-BERT and FLANG-ELECTRA language models, and all the benchmarks in FLUE.

## 2 Related Work

**Pre-trained language models** Language models pre-trained on unlabeled textual data, such as BERT (Devlin et al., 2018), ELMo (Peters et al., 2018) and ROBERTA (Liu et al., 2019), have significantly improved the state-of-the-art in many natural language tasks. Newer models introduce different training objectives: BART (Lewis et al., 2020) uses denoising auto-encoder objective for sequence-to-sequence pre-training; Span-BERT (Joshi et al., 2019) uses a pre-training methodology that predicts spans of text; ELECTRA (Clark et al., 2020) uses token detection for training, where it corrupts some tokens using a generator network and predicts if the tokens are corrupted using a discriminator.

**Masked Language Modeling** Most language models use Masked Language Modeling (MLM) (Devlin et al., 2018) as a training objective. It typically involves randomly masking a percentage of tokens in a text, and using surrounding text to

| Name | Task | Source | Dataset Size | | | Metric | License | Ethical Risks |
|---|---|---|---|---|---|---|---|---|
| | | | Train | Valid | Test | | | |
| FPB | Sentiment Classification | (Malo et al., 2014) | 3488 | 388 | 969 | Accuracy | CC BY-SA 3.0 | Low |
| FiQA SA | Sentiment Analysis | (FiQA)FiQA 2018 | 822 | 117 | 234 | MSE | Public | Low |
| Headline | News Headlines Classification | (Sinha and Khandait, 2020) | 7,989 | 1,141 | 2,282 | Avg F-1 score | CC BY-SA 3.0 | Low |
| NER | Named Entity Recognition | (Alvarado et al., 2015) | 932 | 232 | 302 | F-1 score | CC BY-SA 3.0 | Low |
| FinSBD3 | Structure Boundary Detection | (FinSBD3, 2021)FinWeb-2021 | 460 | 165 | 131 | F-1 score | CC BY-SA 3.0 | Low |
| FiQA QA | Question Answering | (FiQA)FiQA 2018 | 5676 | 631 | 333 | nDCG, MRR | Public | Low |

Table 1: Summary of benchmarks in FLUE. Dataset size denotes the number of samples in the benchmark. Metric denotes the evaluation metric used. Here MSE denotes Mean Squared Error, nDCG denotes Normalized Discounted Cumulative Gain and MRR denotes Mean Reciprocal Rank.

predict the masked tokens. A variety of masking techniques have been used for domain-specific pre-training. While some works (Glass et al., 2020; Sun et al., 2019b) propose rule based masking strategies that work better than random masking, other works (Kang et al., 2020) attempt to find optimal masking policy automatically using techniques such as reinforcement learning.

**Domain-specific Language Models** While the models trained on general English language perform well, domain-specific pre-training can further increase the performance on a particular domain of text (Sun et al., 2019a; Gururangan et al., 2020). For example, BioBERT (Lee et al., 2019) on biomedical domain, ClinicalBERT (Alsentzer et al., 2019) on clinical domain, SciBERT (Beltagy et al., 2019) on scientific publications domain, etc. There have been some works on financial domain as well: previous works by Araci (2019); Yang et al. (2020) directly fine-tune BERT trained on financial corpus for sentiment analysis and question answering tasks respectively. FinBERT (Liu et al., 2020) uses multi-task pre-training to improve performance. The previous works in financial domain rely on basic architectures/ training schemes and do not use finance-specific knowledge. Furthermore, FinBERT is pre-trained with the objective of optimizing performance for sentiment analysis, while we build a generalizable model performing well on a diverse set of tasks. We use and demonstrate that finance specific knowledge and vocabulary can further improve the performance of the model.

**Finance Benchmarks** Wang et al. (2018) created General Language Understudy Evaluation(GLUE), a collection of benchmark tasks for training, evaluating, and analyzing language model designed for non-domain specific tasks. For financial domain, the benchmark suite isn't as exhaustive. Malo et al. (2014) created Financial PhraseBank dataset for Sentiment analysis classification. Maia et al. (2018) created two tasks in (FiQA)FiQA 2018: Task-1 for Sentiment Analysis Regression and Task 2 dataset for Question Answering task in finance. Other datasets include gold news headline datase (Sinha and Khandait, 2020), financial NER (Alvarado et al., 2015) and Structure Boundary Detection (FinSBD3, 2021). Recent financial language models (Araci, 2019; Yang et al., 2020) evaluate their efficacy only on sentiment analysis tasks. We use datasets from existing literature and create a set of heterogeneous benchmark tasks FLUE (Financial Language Understanding Evaluation) for better comprehensive evaluation.

## 3 Benchmarks (FLUE) and Datasets

### 3.1 FLUE

We introduce Financial Language Understanding Evaluation (FLUE), a set of comprehensive benchmarks across 5 financial tasks. The statistics for FLUE are summarized in Table 1 along with the licensing details for public use. All FLUE benchmark datasets have low ethical risks and do not expose any sensitive information of any organization/ individual. Additionally, we have obtained approval for the authors of each dataset for this FLUE benchmark.

#### 3.1.1 Financial Sentiment Analysis

Serving as a fundamental task for textual analysis, this task received a lot of attention in finance domain (Loughran and McDonald, 2011; Garcia, 2013). In our FLUE benchmark, we include both sentiment analysis tasks: regression and classification. For classification, we use Financial PhraseBank dataset (Malo et al., 2014) which provides the sentiment labels annotated by humans for financial news sequences. For regression, we use FiQA 2018 task-1 (Aspect-based financial sentiment analysis)

dataset (Maia et al., 2018), which contains both headlines and microblogs.

#### 3.1.2 News Headline Classification

The financial phrases contain information on multiple dimensions other than the sentiment. Financial news headlines contain important time sensitive information on price changes. To explore our model on those dimensions, we use the Gold news headline dataset created by Sinha and Khandait (2020). The dataset is a collection of 11,412 news headlines, with 9 binary labels.

#### 3.1.3 Named Entity Recognition

Name entity recognition (NER) is key task to analysing any financial text as it can be used along with the Knowledge Graphs to better understand interdependence of different financial entities linked through location, organisation and person. Given a text, NER can identify and classify tokens into specified categories such as person, organisation, location and miscellaneous. We use dataset released by Alvarado et al. (2015) for NER task on financial domain text.

#### 3.1.4 Structure Boundary Detection

Boundary detection of different structure is fundamental challenge in processing text data. Here we employ the dataset shared in the task FinSBD-3 of (FinSBD3, 2021)FinWeb-2021 workshop. The goal of the task is to find the boundaries of different components of text (sentences, lists and list items, including structure elements like footer, header, tables). We chose this dataset as it not only identifies boundaries of sentences but also identifies boundaries of other structural elements.

#### 3.1.5 Question Answering

Question answering system which can answer the finance domain question is essential to any digital assistant. To evaluate our language model's ability on QA task we employ the dataset ("Opinion-based QA over financial data") released in (FiQA) FiQA 2018 open challenge Task 2 (Maia et al., 2018).

### 3.2 Pre-training Datasets

For pre-training, we use a mix of general English language datasets and finance specific datasets. For English, we use BooksCorpus (Zhu et al., 2015) (800M words) and English Wikipedia (2500M words). For the domain specific datasets, we use six publicly available datasets, they are: 1) SEC 10-K and 10-Q financial reports, 2) Earning Conference calls, 3) Analyst Reports, 4) Reuters Financial News, 5) Bloomberg Financial News, and 6) Investopedia. The details for these datasets are summarized in Table 13 and a brief description of each dataset is given in the Appendix Section 7.1.

## 4 Model

For FLANG-BERT, we add financial word and phrase masking, while for FLANG-ELECTRA, we also add a span boundary objective. The addition of financial word and phrasal masking is model agnostic and can be used for any model with a generator.

### 4.1 Financial Word Masking

Previous works (Liu et al., 2020; Yang et al., 2020; Araci, 2019) on financial language modeling use MLM objective for pre-training, which masks some tokens randomly and uses the prediction of those tokens as a training objective. However, there is empirical evidence (Sun et al., 2019b; Kang et al., 2020; Glass et al., 2020) that masking some words strategically which carry more information improves performance on downstream tasks.

Hence, we propose masking financial words preferentially. To this end, we use Investopedia Financial Term Dictionary (Investopedia) to create a comprehensive financial dictionary, which lists the commonly used technical terms in financial markets and literature. We expand our list by adding words/phrases from other financial vocabulary lists available online (Vocabulary.com; MyVocabulary.com; TheStreet).

Our dictionary contains more than 8200 words and phrases. For preferential masking, we mask the single word financial tokens with a probability 30% and randomly mask other tokens with 70% percent probability. Like original BERT pre-training scheme, we mask a cumulative total of 15% of all tokens, such that the total number of tokens being masked in each round is same as the original BERT pre-training approach. Table 10 shows that masking financial terms with a 30% probability gives the lowest perplexity score when pre-training either BERT and ELECTRA with additional vocabulary.

### 4.2 Phrase Masking

Many financial terms are phrases with multiple tokens. It has been shown (Sun et al., 2019b; Joshi et al., 2019) that masking phrases instead of words could leads to better learning of the phrase content.

| Model<br>Metric | FPB<br>Accuracy | FiQA SA<br>MSE | Headline<br>Mean F-1 | NER<br>F-1 | FinSBD3<br>F-1 | FiQA QA<br>nDCG |
|---|---|---|---|---|---|---|
| BERT-base | 0.856 | 0.073 | 0.967 | 0.79 | 0.95 | 0.46 |
| FinBERT (Yang et al., 2020) | 0.872 | 0.070 | 0.968 | 0.80 | 0.89 | 0.42 |
| FLANG-BERT(ours) | **0.912** | **0.054** | **0.972** | **0.83** | **0.96** | **0.51** |
| ELECTRA | 0.881 | 0.066 | 0.966 | 0.78 | 0.94 | 0.52 |
| FLANG-ELECTRA(ours) | **0.919** | **0.034** | **0.98** | **0.82** | **0.97** | **0.55** |

Table 2: Summary of results of our models and baselines on benchmarks. FLANG (Financial Language Model) denotes our final model. Average of 3 seeds was used for each model and benchmark.

| Model | MSE | R2 |
|---|---|---|
| SC-V (Yang et al., 2018) | 0.080 | 0.40 |
| RCNN (Piao and Breslin, 2018) | 0.090 | 0.41 |
| BERT | 0.074 | 0.59 |
| FinBERT | 0.070 | 0.57 |
| FLANG-BERT | **0.052** | **0.67** |
| ELECTRA | 0.046 | 0.72 |
| FLANG-ELECTRA | **0.039** | **0.77** |

Table 3: Resuls on FiQA Sentiment Regression.

| Model | F-1 Scores | |
|---|---|---|
| Multi-token | No | Yes |
| CRFs | 0.83 | |
| BERT | 0.805 | 0.788 |
| FinBERT | 0.795 | 0.800 |
| FLANG-BERT | **0.836** | **0.831** |
| ELECTRA | 0.797 | 0.777 |
| FLANG-ELECTRA | **0.822** | **0.818** |

Table 4: Results on Named Entity Recognition. Yes: Set other tokens in word to same label. CRF result is taken from (Alvarado et al., 2015), but they don't specify that whether they set other tokens in word to same label.

Building on that, we use phrase-based masking in the language model. We perform a two-phase training: in the first phase, we only use word masking to mask single tokens and train the language model; in the second phase, we add phrase masking.

For a financial term of token length $n$, we mask it with a probability of 30%. We replace all tokens in a financial phrase with a single `[MASK]` token. We add all the financial phrases in the model vocabulary and predict the phrase with the usual masked language modeling objective.

### 4.3 Span Boundary Objective

We add the Span Boundary Objective to the loss function along with the MLM loss in the pre-training stage, in addition to the word and the phrasal level masking and the modified vocabulary. Our final loss has three parts:

**Masked Language Modeling Loss** is the Maximum Likelihood Loss of the ELECTRA generator ($G$). We also modify the token masking to randomly mask contiguous spans from a geometric distribution of length L $\sim$ Geo(p), which is skewed towards smaller spans. We follow the results of Joshi et al. (2019) and set p = 0.2.

$$L_{MLM}(x, \theta_G) = E(\sum_{i \in masks} -log(P_G(x_i|x_{masked}))$$

| Model | Accuracy | $\%\Delta MP$ |
|---|---|---|
| BERT | 85.6 | |
| FinBERT | 87.2 | |
| FLANG-BERT | **91.2** | 31.25 |
| ELECTRA | 88.1 | 7.03 |
| w/ AD | 91.1 | 30.47 |
| w/ AD + PFV | 91.4 | 32.81 |
| w/ AD + PFV + SBO | 91.9 | 36.71 |
| w/ AD + PFV + SBO + SCL | **92.1** | **38.28** |

Table 5: Results on Financial Phrase Bank Sentiment Classification Dataset (Malo et al., 2014). Accuracy is given as a percentage. Average of 3 seeds was used for all models. Marginal increase in performance is calculated for FLANG-ELECTRA with respect to FinBERT. FV means using Financial Vocabulary for masking, PFV means using both words and phrases in the financial dictionary for multi-stage masking in the pre-training task, SCL means the use of Supervised Contrastive Learning during the fine-tuning stage.

**Discriminator loss** This loss term is the standard ELECTRA implementation. $L_{Disc}$ penalizes if the discriminator detects a token generated by the generator as *replaced* when it is a *non-corrupt* token or if the token generated by G is *corrupt* and the discriminator detects it as *original*.

**Span Boundary Objective** This term penalizes the low probability of a token being generated given span boundaries (the the representations of tokens present before and after the masked contiguous span). The position of the left boundary token is $x_{start-1}$ and the position of the right boundary token is $x_{end+1}$. By looking at words before and after spans and then trying to generate the tokens in the span, this term helps the model to

| Category | SVM | BERT | FinBERT | FLANG-BERT | ELECTRA | FLANG-ELECTRA | $\%\Delta MP$ |
|---|---|---|---|---|---|---|---|
| Price or Not | **0.965** | 0.955 | 0.956 | 0.960 | 0.951 | 0.964 | 18.18 |
| Price Up | 0.924 | 0.939 | 0.945 | 0.951 | 0.946 | **0.964** | 34.54 |
| Price Constant | 0.715 | 0.980 | 0.978 | 0.981 | 0.977 | **0.987** | 40.90 |
| Price Down | 0.932 | 0.950 | 0.958 | 0.965 | 0.959 | **0.974** | 38.09 |
| Past Price | 0.965 | 0.947 | 0.952 | 0.955 | 0.943 | **0.975** | 47.91 |
| Future Price | 0.732 | 0.987 | 0.985 | **0.988** | 0.984 | **0.988** | 20.00 |
| Past News | - | 0.950 | 0.951 | 0.952 | 0.945 | **0.956** | 10.20 |
| Future News | - | 0.989 | 0.993 | 0.993 | 0.991 | **0.994** | 14.28 |
| Asset Comparison | 0.994 | 0.998 | 0.998 | **0.999** | 0.996 | 0.998 | 0 |
| Mean F-1 Score | 0.890($\bar{7}$) | 0.967 | 0.968 | 0.973 | 0.966 | **0.978** | 31.25 |

Table 6: Results on News Headline Classification. SVM results are taken from (Sinha and Khandait, 2020). All values are F1 scores. FLANG denotes our model. Average of 3 seeds was used for all models. FLANG-ELECTRA also uses Supervised Contrastive Learning while fine-tuning. Marginal increase in performance is calculated for FLANG-ELECTRA with respect to FinBERT.

| Model | F-1 Scores | |
|---|---|---|
| Special tokens | No | Yes |
| BERT | 0.950 | 0.948 |
| FinBERT | 0.872 | 0.890 |
| FLANG-BERT | 0.964 | 0.958 |
| ELECTRA | 0.938 | 0.968 |
| FLANG-ELECTRA | 0.966* | 0.967* |

Table 7: Results on Structure Boundary Detection. *indicates the best model when the combined F1 score of both special tokens is considered. Yes and No are additional special tokens. Average of 3 seeds was used.

build multi-word representations of financial terms that are not captured in our vocabulary.

$$L_{SBO}(x, \theta_G) = E(\sum_{i \in masks} -log(P_G(x_i|y_i)))$$

$$\text{where } y_i = f(x_{start-1}, x_{end+1}, pos_{i-start+1})$$

Here the function $f(\dot{c})$ is the representation function for the $i^{th}$ token in the span and is defined by two feed forward layers:

$$y_i = \text{LayerNorm(GELU}(w_2 * h1)$$

$$\text{where } h_1 = \text{LayerNorm(Gelu}(w_1 * h_0))$$

$$\text{and } h_0 = [x_{start-1}, x_{end+1}, pos_i]$$

Our model is then pre-trained and optimized based on this combined loss function.

$$\text{Total Loss} = L_{MLM}(x, \theta_G) + \lambda_1 L_{SBO}(x, \theta_G)$$

$$+\lambda_2 L_{Disc}(x, \theta_D)$$

| Model | nDCG | MRR | Precision |
|---|---|---|---|
| BERT | 0.46 | 0.42 | 0.35 |
| FinBERT | 0.42 | 0.37 | 0.29 |
| FLANG-BERT | 0.51 | 0.46 | 0.36 |
| SpanBERT + AD + FV + PFV | **0.57** | **0.54** | **0.50** |
| ELECTRA | 0.52 | 0.49 | 0.43 |
| FLANG-ELECTRA | **0.55** | **0.51** | **0.45** |

Table 8: Results on Question Answering benchmark. Average of 3 seeds was used for all models.

### 4.4 Contrastive Loss for Fine-tuning

While most language models are fine-tuned for supervised classification by using cross-entropy loss (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019), we use additional supervised contrastive learning loss for fine-tuning for classification (Gunel et al., 2021). This loss function captures the similarities between examples of the same class and contrasts them with the examples from other classes. Details about Supervised Contrastive Loss are given in Appendix Section 7.3. Here, we only add this loss to the fine-tuning of Financial Phrasebank Dataset and the Headlines Dataset as shown in Tables 5 and 6.

## 5 Experiments

### 5.1 Experiment Setup

All experiments were conducted with PyTorch (Paszke et al., 2019) on NVIDIA V100 GPUs. We initialized each model with their respective pre-trained version on the Huggingface's Transformers library (Wolf et al., 2020). We further pre-trained each model for 4 more epochs on the training data. We used 2 epochs with only single token masking and the later 2 epochs for both word and phrase masking. Using this multi-stage setup gives the lowest model perplexity as shown in Table 11.

We used ELECTRA-base pre-trained model as

our base architecture. ELECTRA corrupts the input by replacing tokens with words sampled from a generator and trains a discriminator model that predicts whether each token in the corrupted input was replaced by a generator sample. This enables it to learn from all input tokens rather than just masked out tokens and is a good fit for our preferential masking approach. We compare our results the following models:

- BERT-base and ELECTRA-base: We use the BERT-base model (Devlin et al., 2018) and the ELECTRA-base model (Clark et al., 2020) from Huggingface (Wolf et al., 2020) and fine-tuned it directly for our tasks.
- finBERT (Yang et al., 2020): We used finBERT model and fine-tune on our tasks.
- **FLANG-BERT (ours)** (Financial LANGuage Model based on BERT): For direct comparison with finBERT, we use our method to train a BERT-base model on our training corpus in a multi-stage manner (Table 11), masking single tokens from financial vocabulary in the first stage and then masking both words and phrases in the second stage.
- ELECTRA w/ AD (Additional Data): The ELECTRA base model pre-trained on our financial training corpus.
- ELECTRA w/ AD + FV (Financial Vocabulary): The ELECTRA Base model is pre-trained on our training corpus, while masking single tokens from financial vocabulary with a higher probability.
- ELECTRA w/ AD + PFV (Phrase Financial Vocabulary). The ELECTRA Base model pre-trained on our training corpus in a multi-stage manner (Table 11), masking only single-word tokens from financial vocabulary in the first stage and masking both words and phrases in the second stage.
- **FLANG-ELECTRA** (Financial LANGuage Model based on ELECTRA): ELECTRA w/ AD + PFV (Phrase Financial Vocabulary) + SBO (Span Boundary Objective). It is pre-trained on our training corpus in the described multi-stage manner with the span boundary and in-filling training objective.
- ELECTRA w/ AD + PFV + SBO + SCL (Contrastive Loss): We use our final language model (FLANG-ELECTRA) but add a contrastive loss term to fine-tune on supervised classification tasks.

### 5.2 Benchmark Results

Summarized results on all benchmarks of our model and baselines are shown in Table 2.

#### 5.2.1 FPB Sentiment Classification

The results of sentiment classification on Financial Phrase Bank sentiment dataset are shown in Table 2. From the accuracy numbers listed in the Table 2, it is evident that FLANG-BERT improves hugely on performance of FinBERT and our final language model(FLANG-ELECTRA) significantly outperforms all the baseline models on the sentiment classification task on the Financial Phrase Bank dataset, achieving state of the art results. Results in Table 5 highlight the importance of each step in our experiment setup described in Section 5.1. As the previous state of art performance on this dataset is already in the higher 80s, we use an additional metric: marginal increase in performance over FinBERT ($\Delta MP$) to demonstrate our techniques. We calculate ($\Delta MP$) as given in equation 1:

$$\Delta MP = \frac{Metric_{Model} - Metric_{FinBERT}}{1 - Metric_{FinBERT}} \quad (1)$$

where the Metric is Accuracy for the Financial Phrasebank Dataset and is F1 score for News Headlines Dataset.

#### 5.2.2 FiQA Sentiment Regression

The results of sentiment regression analysis on the FiQA dataset are shown in Table 3. Evaluation of models is done on two regression evaluation measures Mean Squared Error (MSE) and R Square (R2). Our transformer based architectures outperform conventional techniques like SCV and RCNN. FLANG-BERT model achieves significant improvement on both BERT and finBERT and FLANG-ELECTRA outperforms all models and achieves state of art result for the sentiment regression analysis task on the FIQA dataset.

#### 5.2.3 News Headline Classification

The results of news headline classification for 9 binary classification tasks on Gold headline dataset are shown in Table 6. All the deep learning based language models perform much better than Support Vector Machines. Our ELECTRA-based language model (FLANG-ELECTRA) achieves the highest

| Model<br>Metric | FBP<br>Accuracy | Headline<br>Mean F-1 | NER<br>F-1 | FinSBD3<br>F-1 | FIQA SA<br>MSE | FIQA QA<br>nDCG |
|---|---|---|---|---|---|---|
| BERT | 0.856 | 0.967 | 0.79 | 0.949 | 0.073 | 0.46 |
| BERT + AD | 0.902 | 0.968 | 0.811 | 0.954 | 0.058 | 0.47 |
| BERT + AD + FV + PFV (FLANG-BERT) | 0.912 | 0.972 | **0.834** | 0.962 | 0.054 | 0.51 |
| Distilbert | 0.844 | 0.963 | 0.776 | 0.934 | 0.075 | 0.45 |
| Distilbert + AD | 0.898 | 0.965 | 0.806 | 0.944 | 0.064 | 0.46 |
| Distilbert + AD + FV + PFV | 0.901 | 0.965 | 0.812 | 0.958 | 0.057 | 0.49 |
| SpanBERT | 0.852 | 0.962 | 0.774 | 0.935 | 0.078 | 0.53 |
| SpanBERT + AD | 0.901 | 0.962 | 0.789 | 0.951 | 0.063 | 0.55 |
| SpanBERT + AD + FV + PFV | 0.904 | 0.969 | 0.792 | 0.959 | 0.056 | **0.57** |
| ELECTRA | 0.881 | 0.966 | 0.782 | 0.954 | 0.066 | 0.52 |
| ELECTRA + AD | 0.911 | 0.973 | 0.803 | 0.959 | 0.052 | 0.53 |
| ELECTRA + AD + FV + PFV | 0.914 | 0.977 | 0.825 | 0.962 | 0.038 | 0.55 |
| ELECTRA + AD + FV + PFV + SBO (FLANG-ELECTRA) | **0.919** | **0.978** | 0.816 | **0.967** | **0.034** | 0.56 |

Table 9: Ablation Studies: Average of three seeds were used for each model and benchmark

| *Model Perplexities* | BERT | | ELECTRA | |
|---|---|---|---|---|
| % of Financial Terms Masked | FV | PFV | FV | PFV |
| 10 | 23.02 | 22.88 | 19.10 | 18.96 |
| 20 | 21.45 | 21.30 | 18.44 | 18.42 |
| 30 | **20.29** | **19.53** | **17.87** | **17.52** |
| 40 | 20.80 | 20.11 | 18.67 | 17.98 |

Table 10: Model Perplexities when different percentages of Financial terms are masked. FV means using Financial Vocabulary for masking, PFV means using both words and phrases in the financial dictionary for multi-stage masking in the pre-training task.

| Number of Epochs | | Model Perplexity | |
|---|---|---|---|
| FV | FV + PFV | BERT | ELECTRA |
| 4 | 0 | 20.29 | 17.87 |
| 3 | 1 | 20.11 | 17.82 |
| 2 | 2 | 19.53 | 17.52 |
| 1 | 3 | 20.13 | 17.80 |
| 0 | 4 | 20.05 | 17.69 |

Table 11: Model Perplexities when using multi-stage financial term masking for pre-training. FV means using Financial Vocabulary for masking, PFV means using both words and phrases in the financial dictionary for multi-stage masking.

mean F-1 score compared to other language models. FLANG-BERT performs better than BERT, which again highlights the importance of our setup.

#### 5.2.4 Named Entity Recognition

The results of NER on financial NER dataset provided by (Alvarado et al., 2015) are shown in Table 4. The margin of improvement is more muted in this benchmark. Our models outperform the baselines in a multi-token setting. The multi-token setting refers to all tokens in a word being set to the same label when a word is split into multiple tokens, instead of only labeling the first token and ignoring the rest. Our hypothesis is that when the task doesn't require domain specific knowledge, like NER, pre-training language model on domain specific data does not help.

#### 5.2.5 Structure Boundary Detection

The results of structure boundary detection task on FinSBD3 dataset from (FinSBD3, 2021)FinWeb-2021 are shown in Table 7. In this table, note that the "Special Tokens" setting refers to adding special tokens that are commonly used by pre-trained transformers such as [CLS] to the input. Our models perform similarly or slightly better to baseline architectures. This could be because SBD, like NER, relies more on language cues rather than finance keywords for inference and further gives evidence to the hypothesis that when the task doesn't require domain specific knowledge, one should not get improvement by pre-training a language model on domain specific data. However, our model still performs significantly better than FinBERT.

#### 5.2.6 Question Answering

On Question-Answering, our models outperform the previous works, as shown in Table 8. For evaluation, we compare the following metrics (Michael and Joseph): Precision, nDCG—A higher value means that more relevant documents are retrieved first, and MRR—A higher value means that the first relevant item is retrieved earlier. FLANG-BERT, FLANG-ELECTRA outperform other models on all metrics by a huge margin, but do not outperform SpanBERT pre-trained with Additional Data.

### 5.3 Ablation Studies

We conduct multiple ablation studies to understand the individual impact of our techniques on perfor-

| Model | Perplexity | Size |
|---|---|---|
| BERT-base | 23.66 | 110M |
| FinBERT | 21.11 | 110M |
| FLANG-BERT | 19.53 | 110M |
| Electra | 20.10 | 110M |
| w/ AD | 19.20 | 110M |
| w/ AD + FV | 17.87 | 110M |
| w/ AD + PFV | 17.52 | 110M |
| w/ AD + PFV + SBO | 17.34 | 110M |

Table 12: Comparison of perplexity of our model and baselines. The model size is given in terms of number of parameters, and perplexity is averaged over all sentences in the validation dataset. Average of 3 runs was used for perplexity numbers. Here AD means Additional financial data, FV means using Financial Vocabulary for masking, PFV means using both words and phrases in the financial dictionary for multi-stage masking, and SBO means using the span boundary objective in the pre-training task.

mance. Our studies in Table 10 show that preferentially masking 30% of the financial tokens gives the least perplexity for each model. Furthermore, we find that using single-word financial terminologies in the first two pre-training epochs and multi-word terminologies in the next two gives the lowest perplexity score (Table 11). Table 9 shows that the use of additional data and domain specific preferential masking give substantial increase in performance for our FLUE tasks. Addition of the Span Boundary Objective on the ELECTRA generator gives the best performing model when compared to other similar encoder based architectures like SpanBERT, DistilBERT and BERT. In Table 12, we also show that pre-training models using our methodology gives the lowest perplexity scores when compared to prior baselines. The details for the studies can be found in Table 9 and Appendix Section 7.2.

### 5.4 Discussion

In conclusion, both FLANG-ELECTRA and FLANG-BERT outperform the base architectures (ELECTRA and BERT, respectively). FLANG-BERT also outperforms FinBERT on all the benchmarks, with the same number of parameters. Additionally, on relatively domain-agnostic tasks such as Named Entity Recognition, the improvements are muted. The performance is hugely improved in tasks which utilize finance specific language, such as sentiment analysis, sentence classification and question answering. Overall, the dramatic improvement in most benchmarks suggests that our technique yields state-of-the-art financial language models. We also note that our vocabulary based preferential masking training methodology is both architecture and domain independent and can be generalized to other language models and domains.

## 6 Conclusion

We contribute two language models in the finance domain, which use domain-specific word and phrase masking as a pre-training objective. Additionally, we contribute a comprehensive suite of benchmarks in finance domain across 5 natural language tasks, including new benchmarks using public sources. Our language model outperforms previous language models on all the benchmarks. We will release our models, code and benchmark data on acceptance. We also note that our method is not specific to finance and can be used for any domain-specific language model training.

## Acknowledgements

We would like to thank the anonymous reviewers for their comments. We appreciate the generous support of Azure credits from Microsoft made available for this research via the Georgia institute of Technology Cloud Hub. This work is supported in part by the J.P. Morgan AI Faculty Research Award. Any opinions, findings, and conclusions in this paper are those of the authors only and do not necessarily reflect the views of the sponsors.

## Ethics Statement

We give full credit to the respective authors of each dataset included in our FLUE benchmark and have obtained their permissions for the inclusion of each dataset in FLUE. All FLUE benchmark datasets have low ethical risks and do not expose any sensitive or personal identifiable information. We also obtain explicit permissions to use the datasets given in section 13 for pre-training of the FLANG models from the respective sources.

We understand that training large language models has big carbon-footprint and we have tried to minimize the number of full-scale pre-training runs. The addition of preferential masking and the span boundary objective have minimal computation overhead when compared to pre-training traditional BERT/ELECTRA. We hope that future models work towards lower carbon footprint to reduce the environment costs of pre-training for more sustainable and ethical AI.

## Limitations

Some limitations to our work are: 1) We have not included abstractive generation or summarization tasks in the FLUE benchmark, due to a lack of large, annotated datasets. Future work can be directed towards summarization efforts for the financial domain. 2) We do not include social media data like twitter and reddit in our pre-training step, despite the heavy impact of social media on some financial markets like crypto currencies. This is because of the informal usage of textual data which impedes the formal and syntactical correctness of most financial documents. 3) The models are trained and tested on English tasks and may not perform well on non-English text. The limited availability of non-English domain specific vocabulary makes building multi-lingual FLANG models difficult. 4) While the methodologies presented in this paper can work well for any similarly structured domain like clinical data, it is often difficult to obtain a vocabulary term lists and dictionaries for certain domains. 5) We limit ourselves to using encoder based architectures due to the nature of the popular financial domain specific tasks. Future works can explore the use of other models like GPT3 and T5 for the domain.

## References


Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *ArXiv*, abs/1908.10063.

Paul Asquith, Michael B Mikhail, and Andrea S Au. 2005. Information content of equity analyst reports. *Journal of financial economics*, 75(2):245–282.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Khrystyna Bochkay, Jeffrey Hales, and Sudheer Chava. 2020. Hyperbole or reality? investor response to extreme language in earnings conference calls. *The Accounting Review*, 95(2):31–60.

Robert M Bowen, Angela K Davis, and Dawn A Matsumoto. 2002. Do conference calls affect analysts' forecasts? *The Accounting Review*, 77(2):285–316.

Matthias MM Buehlmaier and Toni M Whited. 2018. Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies*, 31(7):2693–2728.

Brian J Bushee, Dawn A Matsumoto, and Gregory S Miller. 2003. Open versus closed conference calls: the determinants and effects of broadening access to disclosure. *Journal of accounting and economics*, 34(1-3):149–180.

Sudheer Chava, Wendi Du, and Baridhi Malakar. 2020. Do managers walk the talk on environmental and social issues?

Sudheer Chava, Wendi Du, and Nikhil Paradkar. 2019. Buzzwords? *Available at SSRN 3862645*.

Sudheer Chava, Wendi Du, Agam Shah, and Linghang Zeng. 2022. Measuring firm-level inflation exposure: A deep learning approach. *Available at SSRN 4228332*.

Sudheer Chava and Nikhil Paradkar. 2016. December doldrums, investor distraction, and stock market reaction to unscheduled news events. *Available at SSRN 2962476*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

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

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1415–1425.

FinSBD3. 2021. Financial sbd 3. https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-task-finsbd-3.

FiQA. Financial question answering. https://sites.google.com/view/fiqa.

Diego Garcia. 2013. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.

Michael R. Glass, A. Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avirup Sil. 2020. Span selection pre-training for question answering. In *ACL*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. *ArXiv*, abs/2011.01403.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Investopedia. Financial term dictionary from investopedia. https://www.investopedia.com/financial-term-dictionary-4769738.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529.

Minki Kang, Moonsu Han, and Sung Ju Hwang. 2020. Neural mask generator: Learning to generate adaptive word maskings for language model adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6102–6120.

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

M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461.

Feng Li. 2010. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, pages 5–10.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*.

Ekstrand Michael and Konstan Joseph. Rank-aware top-n metrics. https://www.coursera.org/lecture/recommender-metrics/rank-aware-top-n-metrics-Wk98r.

MyVocabulary.com. Business, finance and economics vocabulary word list. https://myvocabulary.com/word-list/business-finance-and-economics-vocabulary/.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Xiao Ding Philippe Remy. 2015. Financial news dataset from bloomberg and reuters. https://github.com/philipperemy/financial-news-dataset.

Guangyuan Piao and John G Breslin. 2018. Financial aspect and sentiment predictions with deep neural networks: an ensemble approach. In *Companion Proceedings of the The Web Conference 2018*, pages 1973–1977.

Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ankur Sinha and Tanmay Khandait. 2020. Impact of news on the commodity market: Dataset and results. *arXiv preprint arXiv:2009.04202*.

Chi Sun, Xipeng Qiu, Yige Xu, and X. Huang. 2019a. How to fine-tune bert for text classification? *ArXiv*, abs/1905.05583.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. Ernie: Enhanced representation through knowledge integration. *ArXiv*, abs/1904.09223.

TheStreet. Financial word dictionary. https://www.thestreet.com/topic/46001/financial-glossary.html.

Vocabulary.com. Personal finance and financial literacy. https://www.vocabulary.com/lists/1504643.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Steve Yang, Jason Rosenfeld, and Jacques Makutonin. 2018. Financial aspect-based sentiment analysis using deep representations. *arXiv preprint arXiv:1808.07931*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

# 7 Appendix

## 7.1 Pre-training datasets

Table 13 summarizes the financial datasets used for pre-training. It also presents the percentage of each dataset sampled in one training epoch. A brief description of each dataset used for pre-training is given below:

### 7.1.1 SEC Financial Reports

Most U.S. public firms are required by the U.S. Securities and Exchange Commission (SEC) to submit annual report (10-K) and quarterly report (10-Q), to provide detailed information about the firm's business, risk factors, and financial performance. 10-K and 10-Q filings were analyzed in (Li, 2010; Loughran and McDonald, 2011; Buehlmaier and Whited, 2018; Chava and Paradkar, 2016). We download the 10-K and 10-Q filings from SEC EDGAR during 1993–2020.

### 7.1.2 Earnings Conference Calls

The earnings conference calls are held by public companies to covey critical corporate information to the investors and analysts (Bushee et al., 2003; Bowen et al., 2002). SeekingAlpha, as a crowd-sourced website in the United States, provides investing information for a large number of public companies and publishes textual transcripts of many earnings conference calls. Bochkay et al. (2020) use the earnings conference call transcripts to analyze the stock market response to the language extremity. (Chava et al., 2019) use BERT to construct emerging technology related discussions in earnings calls and evaluate whether it is just hype. (Chava et al., 2020) employ RoBERTa to extract environmental related discussion in earnings calls and analyze whether managers walk their talk. We collect 151,359 earnings call transcripts from SeekingAlpha from Jan. 2000 to Jul. 2019. (Chava et al., 2022) use BERT to construct a text-based firm-level inflation exposure measure on earning call transcripts.

### 7.1.3 Analyst Reports

Security analysts generate reports related to a firms' future performance after collecting and analyzing the relevant information. Most analyst reports contains earnings forecast, stock recommendation, and stock price target (Asquith et al., 2005). We collect around 201 analyst reports on public firms from LexisNexis. This corpus contains the language the analysts use to disseminate the new information and their interpretation of previous released information to the investors.

### 7.1.4 Reuters Financial News

Financial news corpus is helpful in analyzing the language used in business society. The Thomson Reuters Text Research Collection (TRC2) contains over 1.8M financial news stories during 2008–2009,

| Name | Source | Size | Time Period | %age sampled |
|---|---|---|---|---|
| 10-K | SEC EDGAR | 13660 | 1993-2020 | 8 |
| 10-Q | SEC EDGAR | 36402 | 1993-2020 | 5 |
| Earning Call Transcripts | SeekingAlpha | 151359 | 2007-2019 | 1.5 |
| Financial News | Reuters TRC2 Corpus | 106521 | 2007 | 10 |
| Financial News | Bloomberg Corpus | 387220 | 2009 | 5 |
| Analyst Reports | LexisNexis | 201 | 2017-2020 | 100 |
| Investopedia Articles | Investopedia | 638 | NA | 100 |

Table 13: Summary of financial datasets used for pre-training. Model size denotes the number of samples in the dataset. %age sampled denotes the percentage of each dataset we sampled in a single training epoch.

which is deployed in prior literature (Araci, 2019). We use 10% of this corpus to pre-train our model.

#### 7.1.5 Bloomberg Financial News

Bloomberg disseminates business and market news to the market investors. We obtain the publicly available Bloomberg news articles provided by Philippe Remy (2015), which is used in Ding et al. (2014) to predict the return of Standard & Poor's 500 stock (S&P 500) index.

#### 7.1.6 Investopedia

Investopedia is a financial website which serves as a comprehensive financial dictionary and provides definition and explanation for financial terminologies used in business world. We download the 638 articles for the financial concepts, and use them to pre-train our model. These articles not only provide definitions of financial terms, but also show how they are interrelated to each other.

### 7.2 Ablation Studies

#### 7.2.1 Preferential Masking with Financial Vocabulary

For the first study, we try different configurations while preferentially masking financial terms in the pre-training. Table 10 shows the impact of masking different percentages of Financial Terms on the model perplexity. The perplexities are calculated while keeping the total percentage of masked tokens for all vocabulary at 15 percent. Table 10 shows that masking 30 percent of financial terms gives the least perplexity on the validation set. We also experiment with the multi-stage masking, where in the first stage (first n epochs) we use only the single-word financial tokens and in the second stage (next m epochs) we use both: word and phrasal financial vocabulary masking. Table 11 shows that masking single-word financial vocabulary in the first 2 epochs and masking all financial terms has the lowest perplexity score.

#### 7.2.2 Perplexity on Validation Set

For the second study, we compute perplexity of the language model on the validation set after pre-training. We report the perplexity scores in Table 12. We notice that FLANG-BERT significantly lowers the perplexity on validation set, relative to BERT and FinBERT (Araci, 2019). Despite all models having the same number of parameters, ELECTRA based models show lower perplexity scores. For ablation study, we keep ELECTRA architecture fixed and notice that pre-training with financial data along with general English data lowers perplexity compared to base ELECTRA. Further reduction is seen when using our token masking approach with financial keywords, suggesting that domain specific masking is helpful for domain specific language models. Pre-training with phrase based masking with the span boundary objective in the generator stage results in the best performance, validating the performance of our technique.

#### 7.2.3 FPB Sentiment Classification

For the third study, we fine-tune the models for sentiment analysis on the Financial PhraseBank Dataset (Malo et al., 2014) and report the accuracy in Table 5. We perform a detailed ablation study on ELECTRA architectures with our various techniques. The results suggest that pre-training on financial data improves accuracy from 88.1% to 91.1%, and using a financial vocabulary for token masking further improves the performance to 91.4%. Span boundary objective is even more effective, improving accuracy to $> 91.5\%$. Using contrastive learning for fine-tuning further enables an accuracy of 92.1%, which is significantly higher than previous works.

### 7.3 Supervised Contrastive Loss

Language models are usually fine-tuned (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019) for supervised classification tasks by using cross entropy loss $L_{CE}$:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{C} y_{i,c}\ log\ \hat{y}_{i,c} \tag{2}$$

where $N$ is the number of samples, $C$ is the number of classes, $(x_i, y_i)$ are the sentence and label pairs for sample $i$ and $\hat{y}_{i,c}$ is the model output for probability of sample $i$ having class $c$.

Gunel et al. (2021) showed that using an additional supervised contrastive learning loss $L_{SCL}$ for fine-tuning pre-trained language models improves performance. The loss is meant to capture the similarities between examples of the same class and contrast them with the examples from other classes:

$$L_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i}-1}\sum_{j=1}^{N} \mathbb{1}_{i\neq j}\mathbb{1}_{y_i=y_j}( log\frac{exp(\phi(x_i)\dot{\phi}(x_j))}{\sum_{k=1}^{N}\mathbb{1}_{i\neq k}exp(\phi(x_i)\dot{\phi}(x_k))}) \tag{3}$$

where $N_c$ is the number of samples of class $c$.

Overall loss is given by:

$$L = \lambda L_{CE} + (1-\lambda)L_{SCL} \tag{4}$$

where $\lambda$ is a variable for weighing the two losses.