Master’s Thesis

Deriving Fundamental Credit Scores from Bond Prospectuses

- Chain of Thought Reasoning (kind of) on very long context (defying the limited context lengths of open source LLMs)

For every slide in pp, have possible questions/answers in notes

PP Presentation

Recap what Sebastian did.

My idea: the basics

- What part of Sebastians work I can/will use.
- The purpose is not to analyze earnings figures and financial performance. It’s to find information in text that might be useful to predict if an investment is good or bad.
- Models that understand language
- LLM as a classifier
    - Optimal World: Give LLM a prospectus and LLM outputs all fundamental scores with explanations for each.
    - Not possible because of input length limitations, context window constraints, computational efficiency etc...
    - Instead we break down the problem into smaller chunks.
    - Basic Idea:
        - Section Extraction: Break prospectus into sections relevant to each fundamental score (will need guidance here). I expect this part to take a lot of work to get right.

        - Sebastian looked at “Risk Factors”, “Legal Proceedings”, “Business Summary”, and Forward Looking Statement.
        - I want to cover as many of the fundamental scores as possible. I need rules: e.g. fundamental score 1 is derived by considering sections: x, y, z.
    - Summarize: Use LLM to first summarize sections into key points that it can manage to score and analyze.
        - Different summarization strategies
    - Focuced Prompts: Write prompts specifically related to each fundamental score, using relevant extracted text
        - E.g. (simplified): "Evaluate the following risk factor for a company: [Summarized Risk Factors section]. Rate the severity on a scale of 1 to 5.”
    - Sequential Processing: Run the model to derive each fundamental score.
    - Our basic idea can be run without further training (simple methods that can suprise)
        - Zero-shot learning
        - Few-shot learning
- Basic Idea Expanded: **Use Gold Standard** (actual C4 Fundamental Scores)
    - Question: How many credits have we scored? 500?
    1. Prepare dataset (this will take a lot of work).
        a. Pair-up and label:
            i. section in prospectus relevant to specific fundamental score with the
            ii. corresponding fundamental score
        b. train-test split (80-20)
    2. Once we have the dataset, we can finetune many different LLMs.

a. Select models depending on ressources. I'll start small and scale up.
    i. Remember the perspective. A "small" open source LLM today has 1-3 billions of parameters.
b. Some Examples
    i. **T5 and Flan-T5**: T5 models generally handle input sequences up to about 512 tokens, with Flan-T5 designed for similar ranges. This limit also makes direct processing of long texts challenging without some form of preprocessing like summarization.
    ii. **LLaMa models (7B/13B)**: These have improved capabilities and slightly longer context windows than GPT-3.5 but generally still face constraints when dealing directly with extensive texts.
    iii. **Phi-2**: Good at
    iv. Finance Related:

3. Define fine-tuning task

- Problem: I don't assume that we have a detailed breakdown of each fundamental score pointing to the prospectus. How do we aggregate (while reasoning) subscores into 1 fundamental score?
    - Optimal World: our model learns importance of each sub-elements by knowing just the final score.
    - In reality: aggregation is the best we can do (all of these sub-elements sum to this fundamental score of x). Make model learn that.
        - Be aware of model shortcuts!
        - Tangent: If we need reasoning for each score. We could perform "Direct Preference Optimization" to make the model behave like we want favorizing outputs with reasoning. Fow now the plan was to just output a 1-5 score.

- Problem 2: what if we have a section with 1 very substantial risk (5/5 risk), vs 10 less substantial (2/5) risks.

  $\text{Avg} = \frac{\sum_{i=1}^{n} \text{score}_i}{n}$

  $\text{Weighted Score} = \frac{\sum(\text{score}_i \times \text{weight}_i)}{\sum \text{weight}_i}$

  $\text{Max Score} = max(\text{score}_1, \text{score}_2, \ldots, \text{score}_n).$

- Please think about other challenges

a. **Prompt Design**:

   a. E.g. for a risk factor evaluation (simplified): "Evaluate the following risk factor for a company: [Summarized Risk Factors section]. Rate the severity on a scale of 1 to 5.". Ensure that we use the same seed for reproducibility.

b. **Objective Function**: minimize difference between the model's prediction and actual C4 analyst ratings (MSE, or Categorical Cross-Entropy)

   a. How would loss function look like?

4. Fine Tune

   a. Concider light weight fine-tuning using adapter (freeze weights of foundation model)

      i. instead of tuning all the model's weights, you insert trainable layers "adapters", keep the rest of the model's weights frozen. This is less computationally expensive and you reduce risk of overfitting.

   b. Concider domain adaptive pretraining: e.g. continue pretraining on large finance related corpus. This could adapt the model more closely to financial language.

5. Hyper parameter tuning: Pytorch, WandB

6. Evaluate and Iterate

   a. Evaluate fine-tuned model on test set.

   b. F1 score for classification

        c. Depending on results, iterate: adjust prompts, scale up experiment…

    - At this point we hopefully have models that are better at scoring than basic zero-shot / few-shot learning. We might also see one model outperform the rest.

        - **What is the next step?**

- Modelling approach inspired by Sebastian

> “Predicting time to event. This could, for example, be predicting time till downgrade
> k notches below issued rating, or time till some relative spread change. This is then
> a survival analysis problem, and typically, one would predict this using a Cox proportional hazards regression. This allows us to e.g. give a downgrade probability from issue to t. All three of these approaches are viable options, but the latter in particular has a number of nice properties for us. Namely, it allows us to include all data points in our modelling using the two other approaches, we are forced to exclude names with less than three years of history.” - Sebastian

(Disclaimer: Method and model selection is subject to changes since I’m still following the course that the thesis work will be based on.)

- Challenges
    - Finetuning larger models require GPU.

-

Prompt: "*I already have 500-1000 bonds that were rated (1-5 on the listed parameters) by experts based on their prospectus. These can be used as gold standards. How can I finetune the foundation models and improve their performance?*"

Prompt: "*Also, how can I use the scores calculated for the 4 other parameters we score companies on in a model that predicts the time till downgrade below issued rating, or time till some relative spread ? change? This is a survival analysis problem, and typically, one would predict this using a Cox proportional hazards regression. This allows us to e.g. give a downgrade probability from issue to time t.*"

Context Length Issue

1. **GPT-2/GPT-3.5**: have a token limit (1,024 for GPT-3.5 and smaller for GPT-2)
2.

Summarization is the way.

Motivating questions

Why are corporate bonds more interesting than stocks here?

- It's a contract. Bond prospectus equivalent in stock market?

Find mønstre i alle prospectuses fra downgradede bonds. Dette kan i sig selv være et spændende resultat.

- Eksempel: vi har 100 bonds som blev downgraded indenfor første år. I sektion i prospectus, som vi har vurderet er vigtigst, optræder i 70/100 tilfælde den

samme særlige risiko. Vurder sektionen i bond prospectus med ukendt fremtid der har denne feature som 7/10 risk.

Nicki's Brain

Må jeg fokusere på kun bond prospectuses?

- Giver ikke generelle forskningsresultater når det er så domain-specific
    - fx hvis jeg konkluderer at en specifik LLM metode er bedst til at reason i langt finansielt dokument med specifikt format (bond prospectus). Hvad så med samme metode i andre domains (legal fx)?
    - FINANCE FIRST, NLP SECOND: så tester jeg en måde til at forudsige credit downgrade vs baseline metode såsom ud fra spread.
    - NLP FIRST, FINANCE SECOND: så tester jeg forskellige måder at ekstrahere reasoning konklusion fra langt dokument, fx sammenlign few shot, zero shot, etc. og konkluder hvilke metoder der giver bedst accuracy.

...

Nicolai's Brain

Kan high yield bonds sammenlignes generelt eller bør de inddeles i fx:

- rating, e.g. I assume there's a difference in the "Risk Factors" section of a BB-rated bond and a CC-rated bond.
- årstal, sektor, rating, land,

Hvad ville være et godt udgangspunkt?

- Tag en masse tidligere bond prospectuses og label et prospectus efter om virksomheden beholdt samme rating, eller om de så en upgrade eller downgrade.
    - Efter x quarter, 1 år, 2 år, ...
- hvis jeg tager kigger på writeups fra bonds vi har investeret i referer de så til prospectus? Vil jeg se de vigtige paragraffer have et positivt udtryk (fx ingen tegn på en specifik negativ risiko)

    - Give different sections different weights and optimize the weights for highest performance. Optimize score too (maybe 1-10 rating is worse than 1-3 maybe words are better than letters: very bad, bad, neutral, good, very good)

(Hvad med kortere dokumenter, e.g. earnings reports)

- Meningen med prospectus er at finde information i ordene og ikke i tallene
- Desuden er det til at opstøve alle de muligheder som vi ikke allerede følger/investerer i.
- Der findes intet benchmark til "document understanding" i finansiel sammenhæng ("financial reasoning")

NLP metoder

- Reasoning in System 1 vs System 2 tasks
    - Leverage zero-shot cognitive ability of LLMs discovered in LLMs are zero-shot reasoners
- Few shot learning with the same seed for fairness
    - Find information der kan forklare øget risiko i et bonds der er blevet downgraded
    - If I have 300 bond prospectuses of which 100 were downgraded within a year (2y, 3y or 5y), 100 kept their rating, 100 were upgraded within a year (2y, 3y or 5y) how can I research if certain specific sections in the prospectus in each category explicit some patterns?
        - Credit scoring model has 10 parameters rated 1-10. Get rating on each 10 parameters and use as input in cox regression model.
    - Brug til few shot eller fine tune model med disse sektioner
- Zero-shot / few-shot: zero-shot or few-shot learning models for classification task where labeled data are scarce or expensive to obtain.
- Yann Lecun | Lex Podcast

- LLM limitations

Annotation

Formuler kriterier til annotering (fx ud fra C4 Credit Write-ups eller C4 Scoring model)

Sammenlign annoteringer med partner (fx Jan Sebesta).

- Cohen's Kappa

Augmentation

If we find passages that have significant importance, should we augment, e.g. by paraphrasing, backtranslation etc.?

Explainability of the decisions made. Good investment vs Bad investment: German paper : *This allows domain experts to quickly evaluate (and, if required, easily correct) the decisions for the individual criteria. The resulting human feedback will then be fed back to the model training process as additional training data to incrementally improve the model's predictive quality.*

Datasets

CUAD (Hendrycks et al, 2021c)

Automated methods for extracting key sections or terms that predict credit-rating downgrades.

PubMedQA (Jin et al, 2019): focused on biomedical questions

Reading comprehension and question answering in PubMedQA

Use of "Yes," "No," or "Maybe" to respond to questions based on text abstracts offers a framework for designing similar binary or ternary classification tasks around the probability of credit-rating downgrades.

Adversarial QA (Bartolo et al, 2020):

Explores robustness of models against adversarially crafted questions, which could simulate the complexity and nuance found in bond prospectuses. Developing a model that can withstand adversarial attacks may enhance its ability to accurately assess credit risk from complex legal documents.

Idea GraveYard

- dpo:
    - **use a large scale LLM** to generate demonstrations (responses to prompts that we believe are relevant to credit scoring based on a prospectus)
    - **Create preference pairs**: For each demonstration, you instruct the LLM to generate two different responses. These form a pair of answers from which preferences can be discerned.
    - **Gather preference data**: Each pair of responses is evaluated ot determine which is more effective, accurate, or appropriate for the goal of scoring the credit. This evaluation should be done via expert assessments
    - **Apply DPO**: Direct Preference Optimization is used to process the preference data. DPO works by adjusting the parameters of a generator model so that the model's output aligns more closely with the preferred responses in your data.
    - **Training the generator model**: This model is trained using the outcomes from DPO to consistently generate the type of content that was identified as preferred in the initial demonstrations. Essentially, the generator model learns to emulate the best responses provided by the LLM and validated through preference analysis.
    - Once the model has a solid foundation in the domain-specific requirements captured by our internal data, we introduce external human-preference data (e.g. *Stanford Human Preferences Dataset (SHP) 385K collective human preferences over responses to questions/instructions in 18 domains for training RLHF reward models and NLG evaluation models.*)

- Extend the model using Retrieval Augmented Generation (RAG)