ROBERTA GOES FOR IPO: PROSPECTUS ANALYSIS WITH LANGUAGE MODELS FOR INDIAN INITIAL PUBLIC OFFERINGS

Abhishek Mishra[1] and Yogendra Sisodia[2]

[1]Trust Group, India
[2]Conga, India

ABSTRACT

With the advent of large-scale language models in natural language processing (NLP), extracting valuable information from financial documents has gained popularity among researchers, and deep learning has boosted the development of effective text mining models. Prospectus text mining is very important for the investor community to identify major risk factors and evaluate the usage of the amount to be raised during an IPO. In this paper, we investigate how the recently introduced pre-trained language model Roberta can be adapted for this task. We also introduced prospectus-specific sentence transformers for semantic textual similarity along with a dataset to verify the efficacy of our work.

KEYWORDS

IPO, Prospectus, Large Language Models, Semantic Textual Similarity.

1. INTRODUCTION

An Offer Document refers to the prospectus containing information about the public offering or offer for sale. This document contains all the information an investor needs to make an informed investment decision. The prospectus is analysed by merchant bankers, stockbrokers, and the investor community to identify various risk factors and answer questions such as: what are the risk factors involved? What are the related party transactions? Where will the money be deployed after listing, etc. All the financial statements prior to the IPO and any legal disclosures are available in the final offer document. This prospectus contains all the pertinent information an investor needs to make an investment decision. The final offer document must be made public before the company can list in the Indian securities market.

Natural language processing, with the advent of large-scale language models, has been implemented in a variety of fields, including legal and biomedical [1] and [2]. We intend to apply the same methodology to the IPO.

Our contribution is enumerated below:

- We built a large-scale language model for India's IPO. Our solution based on Roberta will help with a wide range of use cases, such as answering questions, recognizing named entities, and classifying sentences and paragraphs.
- We are also presenting sentence transformers based on our large-scale language model. There are a variety of use cases, including semantic similarity and zero-shot learning.

David C. Wyld et al. (Eds): CST, NLMLT, DMS, CLBD, ITCS, VLSIE - 2022
pp. 45-53, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121905

- We are also making public two datasets:
  - OCR text for 100 prospectuses (PDFs are already in the public domain on the website of the Market Regulator.)
  - One dataset containing pairs of semantically similar sentences was extracted from these prospectuses and annotated by one of the authors, who is a subject matter expert.

2. RELATED WORK

Based on BERT [1], SCIBERT [2] is a pretrained language model for scientific text. SCIBERT was tested on a wide range of scientific domain-specific tasks and datasets. SCIBERT does a lot better than BERT-Base and gets new SOTA results on many of these tasks. LEGAL-BERT [3] is a family of BERT models for the legal domain that achieves state-of-the-art outcomes in many end-tasks. Notably, the performance gains are bigger for the hardest end-tasks (such as multi-label classification in ECHR-CASES and contract header, lease details in CONTRACTS-NER), where domain-specific knowledge is more important. BioBERT [4] is a language representation model that has been pre-trained for biomedical text mining. BioBERT does better than previous models at biomedical text mining tasks like NER, RE, and QA, with only minor changes to the architecture for each task. BioBERT's pre-release version has already been shown to be very good at several biomedical text mining tasks, such as NER for clinical notes. FinBERT is an extension of BERT for the financial domain that was pre-trained on a financial corpus and further fine-tuned for sentiment analysis. The authors achieved state-of-the-art results on both datasets employed by a significant margin. For the classification task, this improved the accuracy of the state-of-the-art by 15%.

The closest things we found that might be related to our domain are FinBert and LegalBert However, on close inspection, we found that FinBert focuses on sentiment analysis and LegalBert focuses more on legal cases and contracts. There isn't.

The most prominent large language model is the BERT model architecture. It is based on a multilayer bidirectional transformer. Instead of the traditional left-to-right language modelling goal, BERT is trained on two tasks: predicting randomly masked tokens and predicting whether two sentences go together or not. There are two primary ways to train a domain-specific language modelling task: a) fine-tuning and b) training from scratch. The key difference is that language model fine tuning begins with a model that has already been trained, whereas training a language model from scratch begins with an untrained, randomly initialised model.

2.1. Fine-tuning the Large Language Model

When refining a language model, a previously pre-trained model (e.g., bert-base, etc.) is retrained on a new unlabeled text corpus (using the original, pre-trained tokenizer). In general, this is advantageous if we intend to use a pre-trained individual for a specific task in which the language employed may be highly technical and/or specialized. The technique was utilised effectively in the SciBERT paper. For computational purposes, we have opted for this method.

2.2. Large Language Model Training from Scratch

A brand-new, randomly initialised model is trained on a massive block of text. This will also improve a tokenizer so that it works best with the data you provide. This comes in especially handy when training a language model for a language that lacks publicly available pre-trained models. However, the computational cost of this method is high.

3. METHODOLOGY

3.1. Dataset Preparation

We downloaded 100 prospectuses from the Market Regulator Website [6]. A prospectus is a very long document with close to 500 pages. The number of pages is 41,697. We used open-source Tesseract OCR [7] to extract text for each document. Text is stored in a JSON file. A prospectus is in Pdf format, so we converted Pdf to images (using Poppler Utilities [8]) and then applied Tesseract. The total number of words is 2,20,60,749.

Figure 1. Corpus Preparation

3.1.1. Prospectus STS

A domain expert author prepared a Semantic Textual Similarity (STS) dataset on 1,598 sentence pairs. The domain expert author was given a random set of sentences (derived from the NLTK sentence tokenizer) and asked to find similar sentences and put them in an STS format. 1,373 pairs are semantically similar. The rest are dissimilar cases.

3.1.2. Prospectus Labels

Also, the domain expert author annotated 18 classes for semantically similar cases. A sentence can have multiple classes, and semantically similar cases have the same classes. These are labels used for identifying various risk factors in a prospectus, e.g., utilisation of funds, operational and currency risk.

Table 1. Semantically Similar Sentences

Sentence 1	**Sentence 2**
This being the first public issue of our corporation, there has been no formal market for the Equity Shares of our Corporation	No assurance can be given regarding an active or sustained trading in the Equity Shares nor regarding the price at which the Equity Shares will be traded after listing.
Investments in equity and equity-related securities involve a degree of risk and investors should not invest any funds in the Offer unless they can afford to take the risk of losing their entire investment.	Investors are advised to read the risk factors carefully before taking an investment decision in the Offer.
Unless the context requires otherwise, the financial information in this Prospectus is derived from the Restated Consolidated Financial Statements of our Corporation comprising	Risk Factors – Significant differences exist between Indian GAAP and other accounting principles, such as U.S. GAAP and IFRS
We have included certain non-GAAP financial measures and certain other selected statistical information related to our business,	non-GAAP financial measures and are significantly different from those of non-insurance companies and may require certain estimates and assumptions in their calculation
Investors may be subject to Indian taxes arising out of the sale of the Equity Shares	unless specifically exempted, capital gains arising from the sale of equity shares held as investments in an Indian company are generally taxable in India

Table 2. Semantically Dissimilar Sentences

Sentence 1	**Sentence 2**
This section of Indian society is characterized by low levels of financial literacy and technology use, lack of financial	judicial precedent may be time consuming as well as costly for us to resolve and may impact the viability of our current business.
Except as disclosed in chapter titled “Financial Statements” beginning on page 1494of this Draft Red Herring .	Stringent quality control is followed during the production process by the quality control department by
Credit risk is the risk of financial loss to our Company if a customer or counterparty to a financial instrument fails to meet	Our total expenses marginally increased by 0.49% to = 14,834.82 million in Fiscal 2020 from % 14,762.64 million
Revenue from contracts with customers is recognised upon transfer of control of promised goods/ services	Except as stated in the chapter titled “Capital Structure” beginning on page 63 of this Red Herring Prospectus
We have a wide variety of 18 different vegetarian and non-vegetarian burgers covering both value and premium	it shall provide reasonable assistance to our Company and the BRLMs in the taking of all steps

Table 3. Class labels

Class	**No of Examples**	**Example Sentence**
Operational Risk	6	human and systems errors when executing complex and high-volume transactions
Intellectual Property	100	We cannot ensure that our intellectual property is protected from copy or use by others, including our competitors, and intellectual property

		infringement actions may be brought against us
Employee Reservation Portion	78	DISCOUNT OF ₹45 PER EQUITY SHARE WAS OFFERED TO THE RETAIL INDIVIDUAL BIDDERS BIDDING IN THE RETAIL PORTION AND THE ELIGIBLE EMPLOYEES BIDDING IN THE EMPLOYEE RESERVATION PORTION
Remuneration	122	Our Directors, Key Managerial Personnel and the Promoter have interests in us other than reimbursement of expenses incurred or normal remuneration or benefits
Currency Risk	14	Fluctuations in the exchange rate between the Rupee and other currencies could have an adverse effect
Utilisation of funds	190	Monitoring Utilization of Funds
Liquidity Risk	4	Our investment portfolio is subject to liquidity risk, which could adversely affect its realizable value
environmental social and governance	12	We continue to undertake various initiatives towards this, including alleviating poverty, pursuing inclusive growth, promoting gender equality, promoting good health, reducing our carbon footprint through consumption rationalisation and using eco-friendly technology
Credit Risk	58	We are subject to the credit risk of the issuers whose debt securities we hold
Risk Disclosure	374	This being the first public issue of our corporation, there has been no formal market for the Equity Shares of our Corporation
Remuneration	120	None of our Directors are entitled to remuneration from our Subsidiaries or Associates
Offer	530	The determination of the Price Band is based on various factors and assumptions.
Market Risk	118	Fluctuations in the exchange rate between the Rupee and other currencies could have an adverse effect
Litigations	384	To material litigation in (iv) above, our Board has considered and adopted the following policy on materiality with regard to outstanding litigation
Investor Taxation	66	Investors may be subject to Indian taxes arising out of the sale of the Equity Shares
Related Party Transactions	192	A summary of related party transactions entered into by our Company with related parties as at and for the nine months ended December
Financial Statements	386	Unless the context requires otherwise, the financial information in this Prospectus is derived from the Restated Consolidated Financial Statements of our Corporation
Audit	54	The Statutory Auditor to the Offer has included certain matters of emphasis in its examination report

As part of next steps Authors want to do more deep work with these classes.

3.2. Methods

3.2.1. Fine-Tuning Roberta

In BERT literature, there are two training objectives: Masked Language Model (MLM) and Next Sentence Forecast (NSP). In MLM random subset of the tokens in the input sequence is selected and replaced with the token special [MASK] Cross-entropy loss in predicting masked tokens is the MLM objective. 15% of the input tokens are chosen uniformly for potential replacement by BERT. 80% of the tokens are replaced with [MASK], 10% are left alone, and 10% are replaced with a randomly selected vocabulary token.

NSP is a binary classification loss for predicting whether two segments in the original text follow one another. The creation of positive examples involves selecting consecutive sentences from the text corpus. Negative examples are created by combining sections from various documents. Positivity and negativity are sampled with equal likelihood.
The purpose of the NSP objective was to improve performance on downstream tasks, such as NLI, that require reasoning about the relationships between pairs of sentences.

When pretraining BERT models, the Roberta [9] Team evaluates several design decisions with great care. They found that performance can be greatly improved by training the model longer, in larger batches, with more data, by removing the goal of predicting the next sentence, by training on longer sequences, and by changing the masking pattern on the training data in real time.

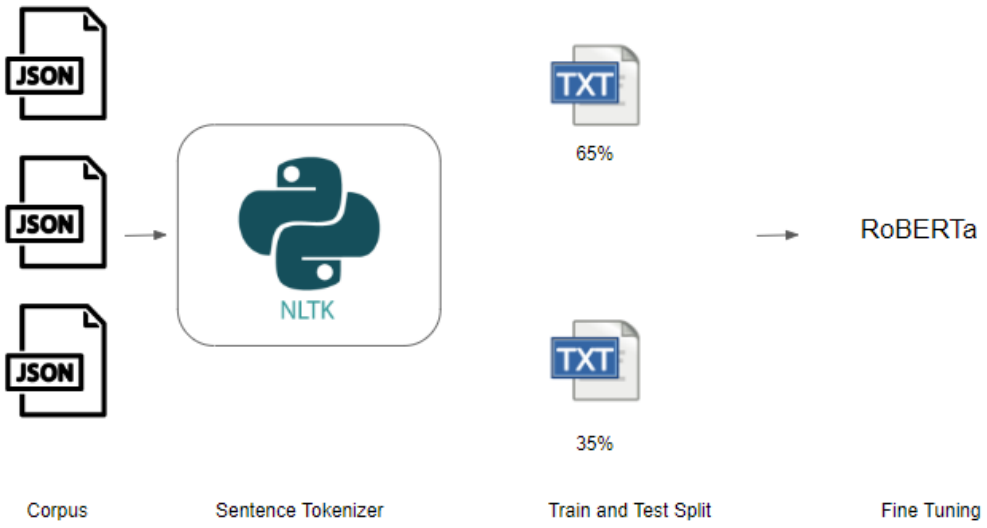

Figure 2. Data Preparation for Roberta Fine-Tuning

We divided the data into 65 and 35% for training and testing. Roberta is fed with sentences derived from NLTK's sentence tokenizer [10]. We fine-tuned our model for 10 epochs. Result is shown in next section.

3.2.2. Sentence Transformer for Semantic Textual Similarity

We used TSDAE-Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning [11] to further get sentence transformer from our pre-trained Roberta TSDAE is a robust method for domain adaptation and pre-training sentence embeddings.

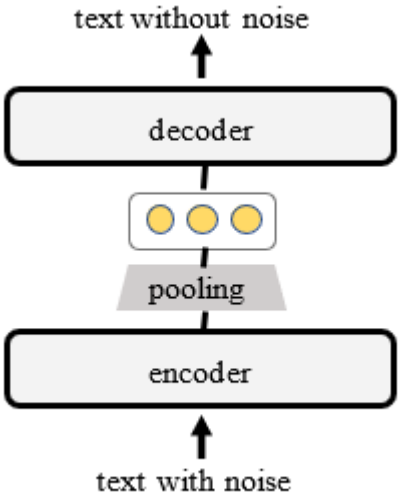

Figure 3. TSDAE Architecture

TSDAE trains sentence embeddings by introducing a specific type of noise (e.g., deleting or exchanging words) into input sentences, encoding the damaged sentences into fixed-size vectors, and then reconstructing the vectors into the original input. The formal training objective is:

$$
\begin{aligned}
(\theta) &= \mathbb{E}_{x\sim D}[\log P_\theta(x|\tilde{x})] \\
&= \mathbb{E}_{x\sim D}[\sum_{t=1}^{l} \log P_\theta(x_t|\tilde{x})] \\
&= \mathbb{E}_{x\sim D}[\sum_{t=1}^{l} \log \frac{\exp(h_t^T e_t)}{\sum_{i=1}^{N} \exp(h_t^T e_i)}]
\end{aligned}
$$

where D is the text corpus, $x = x_1x_2 \cdots x_l$ is the input text training sentence with l no of tokens, $\tilde{x}$ is the equivalent broken sentence, e_t is the word embedding of x_t, N is the vocabulary size and h_t is the hidden state at t decoding step.

All resources for data are available at this link:
https://github.com/scholarly360/ProspectusRoberta

4. RESULTS

4.1. Fine-Tuning Roberta

Perplexity measures, given a model and an input text sequence, the likelihood that the model will generate the input text sequence. It can be used as a metric to evaluate how well the model has learned the distribution of the text it was trained on for the language generation task. Our perplexity on the test dataset was 2.7935, which is a very decent and attainable score.

4.2. Semantic Textual Similarity

The Spearman and Pearson correlation coefficients are normally used for the evaluation of STS datasets. The main difference between the Pearson and Spearman correlation coefficients is that the Pearson value assumes that the two variables are related in a linear way, while the Spearman value also considers monotonic relationships. Table 1 shows our Roberta-based sentence transformer performed better compared to other state-of-the-art sentence transformers [12].

Table 1. Prospectus STS Evaluation Results

Model	**Pearson**	**Spearman**
multi-qa-mpnet-base-dot-v1	0.7228	0.597
all-mpnet-base-v2	0.7369	0.5939
Our Sentence Transformer	0.7859	0.6023

These results show that domain adaption for prospectus is working well. The authors want to further define more problem statements, collect more data, and train larger models.

5. CONCLUSIONS

We have released a Large Language Model based on Roberta for analysing prospectuses. We also put out a Roberta-based sentence transformer that was trained with TSDAE and did a better job on an STS data set with text derived from the Prospectus. This will help investors and the merchant bank community to explore prospectuses in a more automated way, thus saving time.

In the future, we want to explore more with additional use cases such as sentence classification (Zero Shot and Few Shot Classifiers) and token classification. Also, we want to tag our data with the help of multiple annotators so that we can make as accurate a copy as possible of real-world problems.

ACKNOWLEDGEMENTS

The authors would like to thank everyone in their respective organisations for fully supporting them.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee Kristina, Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” 2018. [Online]. Available: arXiv:1810.04805.

[2] Iz Beltagy, Kyle Lo, Arman Cohan, “SciBERT: A Pretrained Language Model for Scientific Text” 2019. [Online]. Available: arXiv:1903.10676.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, Ion Androutsopoulos, “LEGAL-BERT: The Muppets straight out of Law School” 2020. [Online]. Available: arXiv:2010.02559.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang,” BioBERT: a pre-trained biomedical language representation model for biomedical text mining” 2018. [Online]. Available: arXiv:1901.08746,

[5] Dogu Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.”, 2019. [Online]. Available: arXiv:1908.10063,

[6] Public Issues, SEBI, 2022. [Online]. Available: https://www.sebi.gov.in/sebiweb/home/HomeAction.do?doListing=yes&sid=3&ssid=15&smid=12.
[7] Tesseract Documentation, Tesseract. 2022. [Online]. Available: https://tesseract-ocr.github.io/tessapi/4.0.0/.
[8] Poppler, a PDF rendering library, Poppler. 2022. [Online]. Available: https://github.com/freedesktop/poppler/.
[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov," RoBERTa: A Robustly Optimized BERT Pretraining Approach" 2019. [Online]. Available: arXiv:1907.11692.
[10] nltk.tokenize package, NLTK, 2022. [Online]. Available: https://www.nltk.org/api/nltk.tokenize.html.
[11] Kexin Wang, Nils Reimers, Iryna Gurevych, "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning" 2021. [Online]. Available: arXiv:2104.06979.
[12] Sentence Transformer Pretrained Models, Sentence Transformer. 2022. [Online]. Available: https://www.sbert.net/docs/pretrained_models.html.

AUTHORS

As Deputy Vice President at Trust Capital, Abhishek Mishra oversees new issues, private placements, and secondary sales in the Indian Fixed Income markets.

Yogendra Sisodia is Director, Machine Learning at Conga. Some current areas of research interest are semi-supervised learning, adversarial machine learning, and deep learning applications in computer vision.