# Using Natural Language Processing to Enhance Understandability of Financial Texts

Sohom Ghosh and Sudip Kumar Naskar

November 9, 2022

# Using Natural Language Processing to Enhance Understandability of Financial Texts

Sohom Ghosh*
Jadavpur University
Kolkata, West Bengal, India
sohom1ghosh@gmail.com

Sudip Kumar Naskar
Jadavpur University
Kolkata, West Bengal, India
sudip.naskar@gmail.com

## ABSTRACT

Dealing with money has always been one of the basic skills one needs to live a comfortable life. However, financial literacy rates across the nations are extremely low. Furthermore, over the years the returns from traditional investment avenues like bank fixed deposits (FD), real estate, etc. have been diminishing. This entices new-age investors to trade and reap profits from the ever-growing stock markets. Nevertheless, in reality, only a handful of active traders are able to earn more than the FD rates. This is due to the lack of financial knowledge. The presence of complex concepts and jargons further reduces comprehensibility. In this paper, we present how financial texts can be demystified using Natural Language Processing (NLP). It consists of neural-based readability assessment and hypernym extraction tools to improve the readability of financial texts. Other modules include financial domain specific systems for automated claim detection, sustainability assessment, etc.

## CCS CONCEPTS

• **Applied computing** → **Economics**; • **Information systems** → **Clustering and classification**; **Summarization**; • **Computing methodologies** → *Language resources*; *Lexical semantics*; *Information extraction*; Ensemble methods.

## KEYWORDS

financial text processing, natural language processing, readability, hypernym detection, claim detection

## 1 INTRODUCTION

Financial Literacy is understanding how to make the best use of the money one possesses. It leads to financial freedom and improves the overall quality of life. This financial well-being in turn results in economic prosperity of the nation. A survey by National Centre for Financial Education revealed that only 27% of Indians are financially literate.[1] Due to lack of awareness, people tend to invest only in conventional avenues like gold, FDs etc.[2] Since return of investment from these traditional avenues have been constantly reducing, it is essential to enlighten the general public with the knowledge of financial markets and products like equities, bonds, mutual funds, etc.[3] It is equally noteworthy that only a few active traders could earn more than the FD rates over the last 3 years.[4] Thus, apart from improving overall Financial Literacy, it is also essential to equip investors with data-driven tools for making trade-related decisions[5]. In this paper, we describe the four NLP based modules which we have developed so far (Ref: Figure 1).

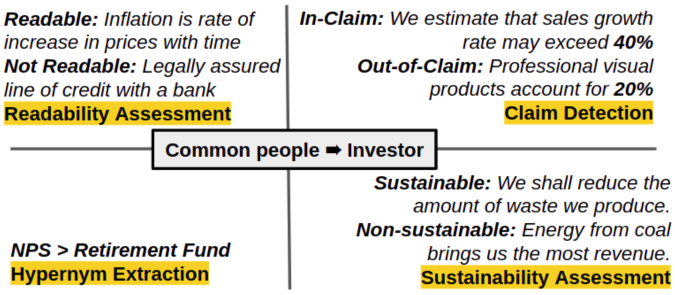
Readable: Inflation is rate of increase in prices with time
Not Readable: Legally assured line of credit with a bank
Readability Assessment
In-Claim: We estimate that sales growth rate may exceed 40%
Out-of-Claim: Professional visual products account for 20%
Claim Detection
Common people ➡ Investor
NPS > Retirement Fund
Hypernym Extraction
Sustainable: We shall reduce the amount of waste we produce.
Non-sustainable: Energy from coal brings us the most revenue.
Sustainability Assessment

**Figure 1: NLP based solutions for the Financial domain**

## 2 MODULES

The first two modules (2.1, 2.2) are aimed at improving Financial Literacy. The remaining two modules (2.3, 2.4) are positioned to empower investors in making informed choices.

### 2.1 Readability Assessment (RA)

We showed that formula-based readability scores (like Flesch Reading Index, Automated Readability Index, etc.) are not suitable for the financial domain and created a Financial Readability Assessment Dataset (FinRAD) [12]. It consists of definitions of financial terms and assigned readability scores. We fine-tuned a FinBERT model [1] and developed the tool FinRead [11] to classify definitions of financial terms as readable or not. In future, we want to develop a tool for simplifying financial text to improve readability.

### 2.2 Hypernym Extraction (HE)

Hypernyms (i.e. generic forms) help in understanding the concepts better. We proposed a novel method to select negative samples [2] and fine-tuned the FinBERT [1] model using sentence transformers (SBERT) [16] to identify hypernyms for a financial term. In future,

[*] Sohom Ghosh is a doctoral student at the Department of Computer Science and Engineering, Jadavpur University, Kolkata.

This manuscript is the authors' pre-print version. It has been accepted at the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), 2023).
https://doi.org/10.1145/3570991.3571051

[1] https://www.financialexpress.com/market/only-27-indians-are-financially-literate-sebis-garg/2134842/

[2] https://indianexpress.com/article/business/market/less-than-1-of-rural-households-invest-in-stocks-sebi-survey-4601264/

[3] https://economictimes.indiatimes.com/markets/stocks/news/financial-education-and-its-importance-in-making-investing-accessible-across-india/articleshow/91792025.cms

[4] https://economictimes.indiatimes.com/markets/stocks/news/no-easy-money-less-than-1-active-traders-beat-bank-fds/articleshow/88656009.cms

[5] https://www.analyticsinsight.net/leveraging-artificial-intelligence-to-simplify-financial-knowledge/

we want to develop a method for automatically extracting an ontology from financial documents. This will help people understand complex concepts by simply looking at the hierarchy.

### 2.3 Sustainability Assessment (SA)

Investors are now looking for Environmental, Social, and Governance compliant sustainable avenues for investment[6]. Understanding the sustainability aspect of the policies of an organization by going through its reports is tedious and challenging. We trained a RoBERTa [14] model [9] to determine the sustainability aspects present in financial texts. This will aid socially responsible investors in decision-making.

### 2.4 Claim Detection (CD)

Many times executives of organizations make false claims to allure investors. Thus, it is essential to empower investors so that they can distinguish facts from claims. We developed tools (FiNCAT [6] & FiNCAT-2 [5]) which use a Logistic Regression based classifier over context-based BERT [3] embeddings to detect whether numerals present in financial earnings conference calls are claims or not. In [8] and [4], we presented several other approaches for this task.

### 2.5 Other Modules

We re-use some of the existing state-of-the-art systems for sentiment analysis[7] [13] and summarization [15] of financial texts.

## 3 RESULTS AND DISCUSSIONS

We combined all the modules and created the *Financial Language Understandability Enhancement Toolkit* (**FLUEnT**)[8,9,10] [10]Table 1 presents the performance of each of the models described above. Extracting causality [7], examining the effect of social media posts on stock prices are other directions for future works. The work can be extended by developing vernacular systems for the people in general. We will focus on demystifying other kinds of financial texts using NLP.

| System | Model | Task | Performance |
|---|---|---|---|
| RA | FinBERT | CLS | AUROC: 0.9927 |
| HE | SBERT+FinBERT | IR | Accuracy: 0.9170 |
| SA | RoBERTa | CLS | Accuracy: 0.9317 |
| CD | BERT | CLS | Macro-F1: 0.8238 |

**Table 1: Results (CLS = Classification, IR = Info. Retrieval)**

## REFERENCES

[1] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063 [cs.CL] https://arxiv.org/abs/1908.10063

[2] Ankush Chopra and Sohom Ghosh. 2021. Term Expansion and FinBERT fine-tuning for Hypernym and Synonym Ranking of Financial Terms. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2021)*. -, Online, 46–51. https://aclanthology.org/2021.finnlp-1.8

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[4] Sohom Ghosh and Sudip Kumar Naskar. 2022. Detecting context-based in-claim numerals in Financial Earnings Conference Calls. *International Journal of Information Technology* – (2022). https://doi.org/10.1007/s41870-022-00952-7

[5] Sohom Ghosh and Sudip Kumar Naskar. 2022. FiNCAT-2: An enhanced Financial Numeral Claim Analysis Tool. *Software Impacts* 12 (2022), 100288. https://doi.org/10.1016/j.simpa.2022.100288

[6] Sohom Ghosh and Sudip Kumar Naskar. 2022. FiNCAT: Financial Numeral Claim Analysis Tool. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)* (Virtual Event, Lyon, France). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3487553.3524635

[7] Sohom Ghosh and Sudip Kumar Naskar. 2022. LIPI at FinCausal 2022: Mining Causes and Effects from Financial Texts. In *Proceedings of the The 4th Financial Narrative Processing Workshop (FNP@LREC2022)*. European Language Resources Association, Marseille, France, 130–132. https://aclanthology.org/2022.fnp-1.21

[8] Sohom Ghosh and Sudip Kumar Naskar. 2022. Lipi at the ntcir-16 finnum-3 task: ensembling transformer based models to detect in-claim numerals in financial conversations. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*. NII, Tokyo, Japan, 92–94. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/02-NTCIR16-FINNUM-GhoshS.pdf

[9] Sohom Ghosh and Sudip Kumar Naskar. 2022. Ranking Environment, Social And Governance Related Concepts And Assessing Sustainability Aspect Of Financial Texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI-ECAI 2022)*. -, Vienna, Austria, 87–92. https://mx.nthu.edu.tw/~chungchichen/FinNLP2022_IJCAI/14.pdf

[10] Sohom Ghosh and Sudip Kumar Naskar. 2023. FLUEnT: Financial Language Understandability Enhancement Toolkit. In *6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (CODS-COMAD 2023)* (Mumbai, India). Association for Computing Machinery, New York, NY, USA, In press. https://doi.org/10.1145/3570991.3571067

[11] Sohom Ghosh, Shovon Sengupta, Sudip Naskar, and Sunny Kumar Singh. 2021. FinRead: A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. NLP Association of India, NIT, Silchar, India, 658–659. https://aclanthology.org/2021.icon-main.81

[12] Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, and Sunny Kumar Singh. 2022. FinRAD: Financial Readability Assessment Dataset - 13,000+ Definitions of Financial Terms for Measuring Readability. In *Proceedings of the The 4th Financial Narrative Processing Workshop (FNP@LREC2022)*. European Language Resources Association, Marseille, France, 1–9. http://lrec-conf.org/proceedings/lrec2022/workshops/FNP/pdf/2022.fnp-1.1.pdf

[13] Allen Huang, Hui Wang, and Yi Yang. 2020. FinBERT—A Large Language Model Approach to Extracting Information from Financial Text. http://dx.doi.org/10.2139/ssrn.3910214

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

[15] Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. 2021. Towards Human-Centered Summarization: A Case Study on Financial News. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 21–27. https://www.aclweb.org/anthology/2021.hcinlp-1.4

[16] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[6] http://bwdisrupt.businessworld.in/article/Sustainable-Investing-To-Surge-To-125-B-In-India-By-2026-Report/09-06-2022-432078/

[7] https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

[8] **Google Colab link:** https://colab.research.google.com/drive/1-KBBKByCU2bkyAUDwW-h6QCSqWI8z127?usp=sharing

[9] **Video:** https://youtu.be/Bp8Ij5GQ59I

[10] **Demo:** https://huggingface.co/spaces/sohomghosh/FLUEnT