Applying Natural Language Processing on Financial Texts

Sohom Ghosh, Jadavpur University, Kolkata, India
sohom1ghosh@gmail.com sohomghosh.github.io

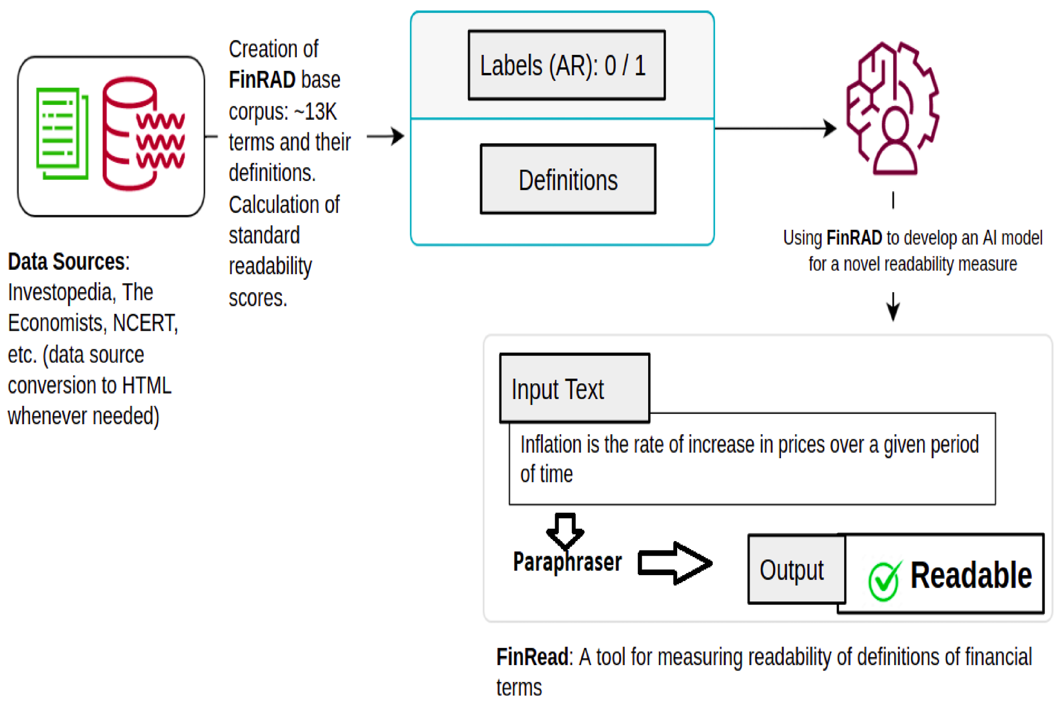
Summary

Humans strive for a better quality of life, which is often facilitated by financial stability. However, several obstacles hinder individuals' progress towards financial prosperity, including insufficient financial literacy, escalating wealth inequality, and the proliferation of misleading information on social media. We explore four key areas where Natural Language Processing (NLP) can contribute to enhancing financial literacy, reducing wealth disparities, ensuring a sustainable future, and fostering economic prosperity. These areas are: Inclusive Investing, Enhanced Investing, Impactful (Green) Investing, and Informed Investing. Additionally, we focus on catering specifically to the Indian market (Indic Investing) and provide various resources to improve the comprehensibility of financial texts. Inclusive Investing focuses on increasing the readability and accessibility of financial texts. Improved Investing aims to streamline the investor's journey by offering hypernyms and relationships between entities. Impactful Investing emphasizes sustainable pathways. Informed Investing involves eliminating financial misinformation from social media, such as assessing the credibility of posts by executives and identifying false or exaggerated claims. In most instances, we demonstrate the effectiveness of our methods by comparing them to existing state-of-the-art techniques.

Inclusive Investing

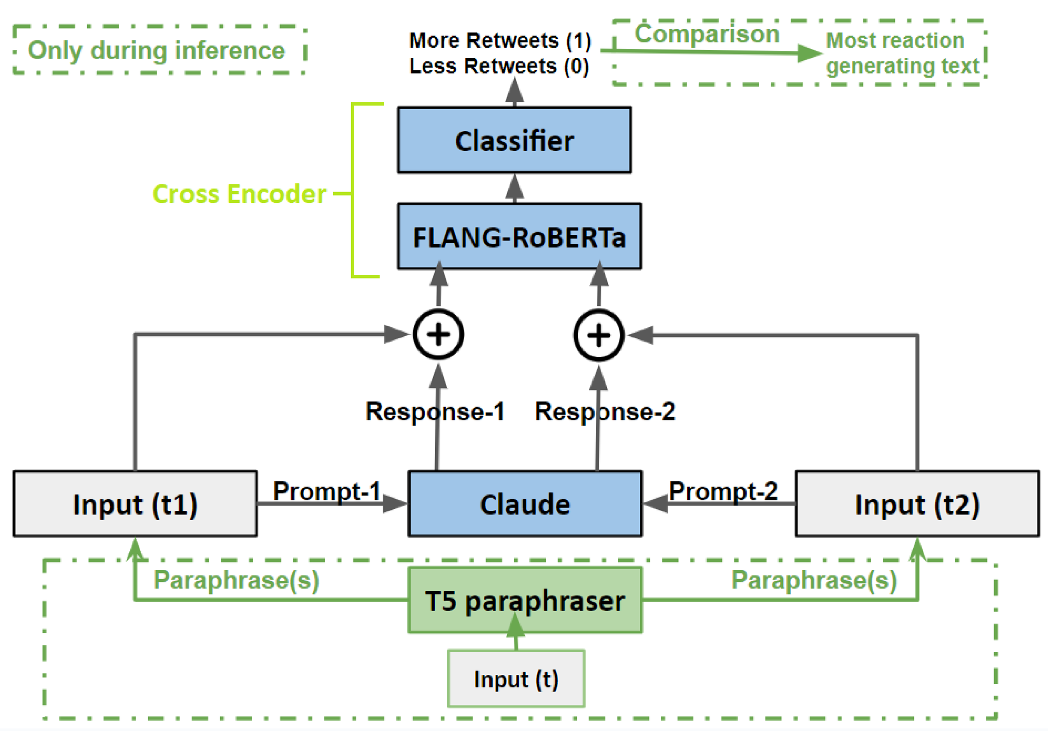
Task-1: Given a financial text (FT), we want to assess its readability and simplify it.

FNP@LREC-2022
ICON-2021



Task-2: Given two FTs, we want to assess which one would to reach more people

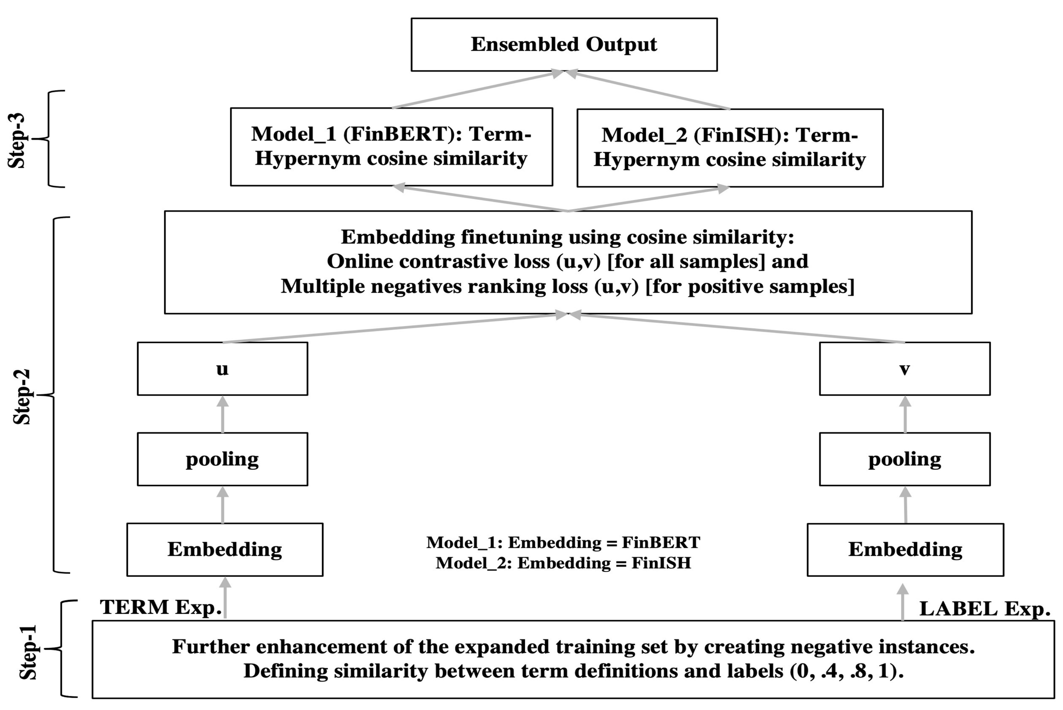
The Web Conf (WWW-2024)



Improved Investing

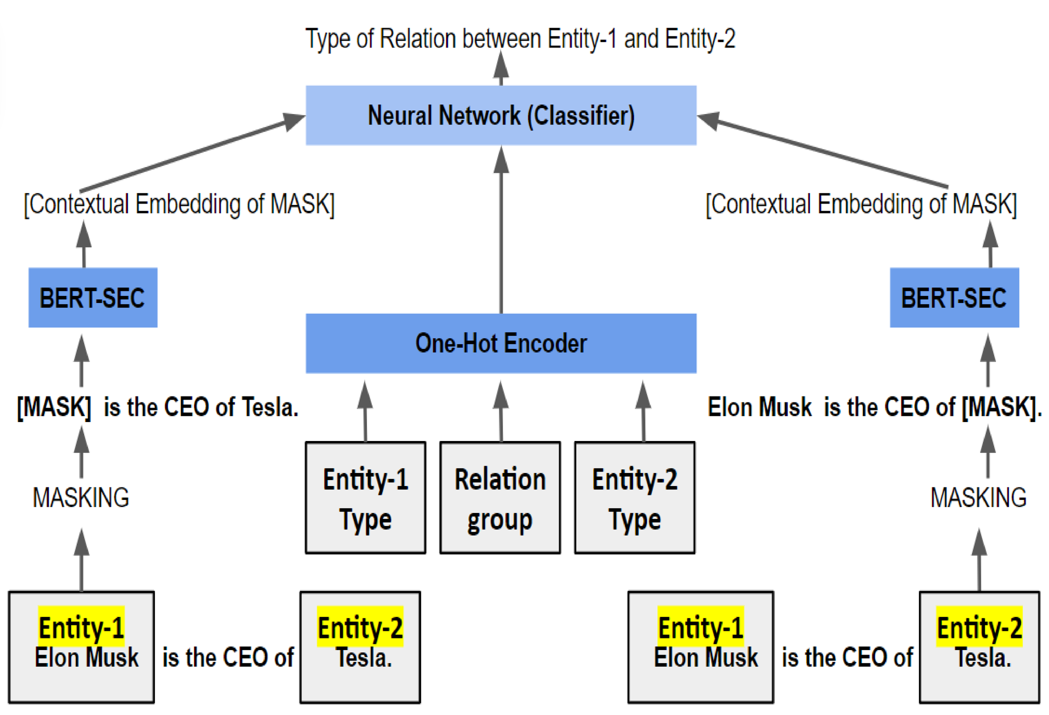
Task-3: Given a financial jargon in a FT, we would like to retrieve its hypernym

FinNLP@IJCAI-2021
SNCS Springer



Task-4: Given two entities in a FT, we would like to determine the relationship between them.

FIRE-2023



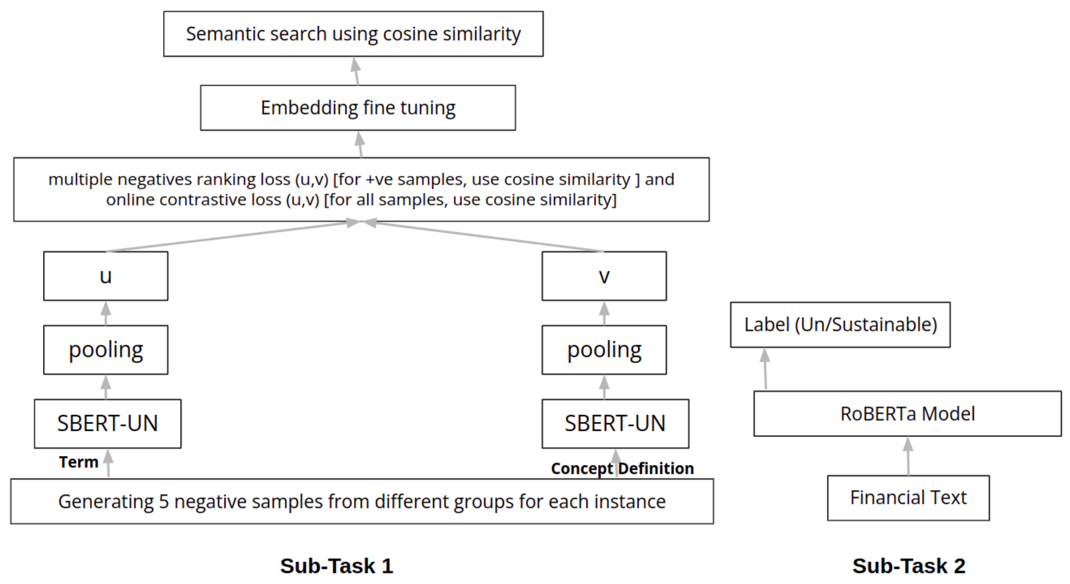
Impactful Investing

Task-5: Classify a FT as Sustainable / Unsustainable

FinNLP@IJCAI-ECAI 2022

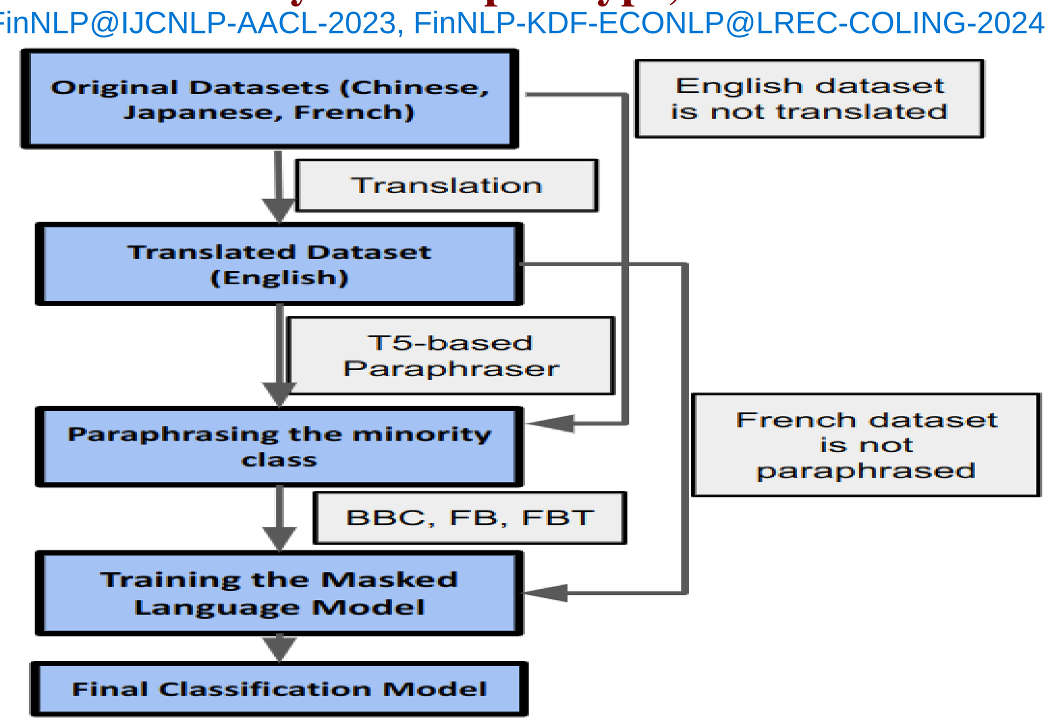
Task-6: Detect ESG Issues from FTs in English

FinNLP@IJCAI-ECAI 2022



Task-7: Identify ESG impact type, duration of FTs

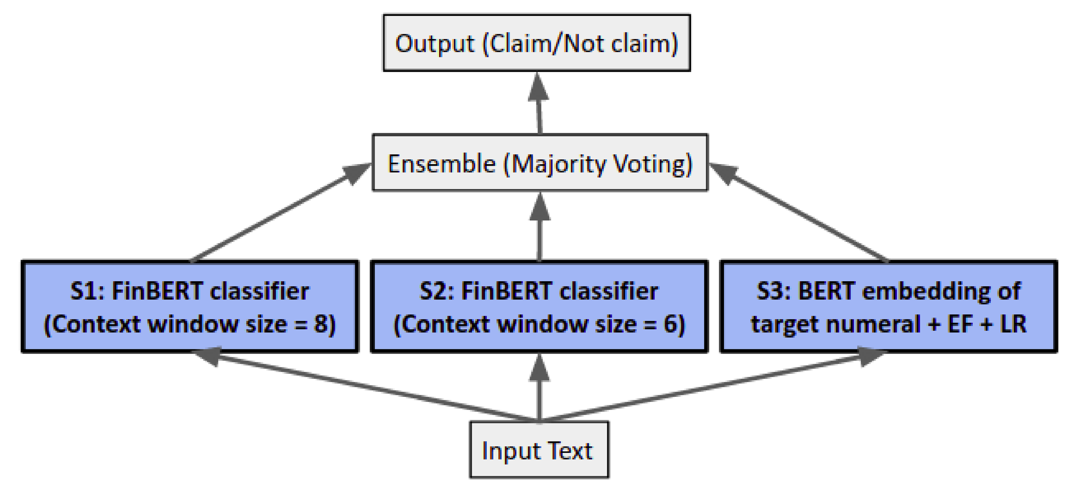
FinNLP@IJCNLP-AACL-2023, FinNLP-KDF-ECONLP@LREC-COLING-2024



Informed Investing

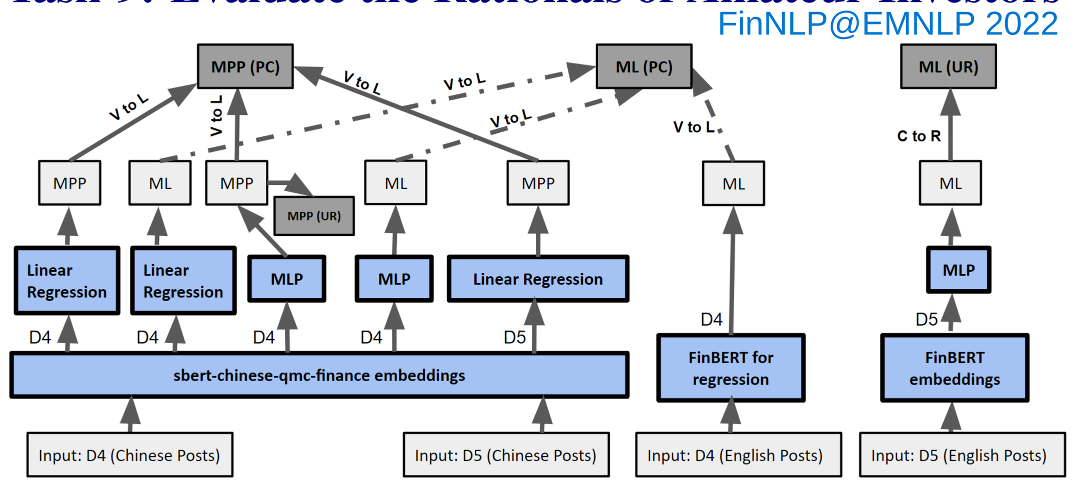
Task-8: Detect exaggerated and in-claim numerals

NTCIR-16, FinWeb@WWW-2022, Software Impacts (Elsevier)



Task-9: Evaluate the Rationals of Amateur Investors

FinNLP@EMNLP 2022



Task-10: Evaluate the trustworthiness of Social Media Posts by Executives on Stock Prices

FIRE-2022

Task-11: Fine-grained Argument Understanding

NTCIR-17

Indic Investing

Task-12: Financial Argument Analysis in Bengali

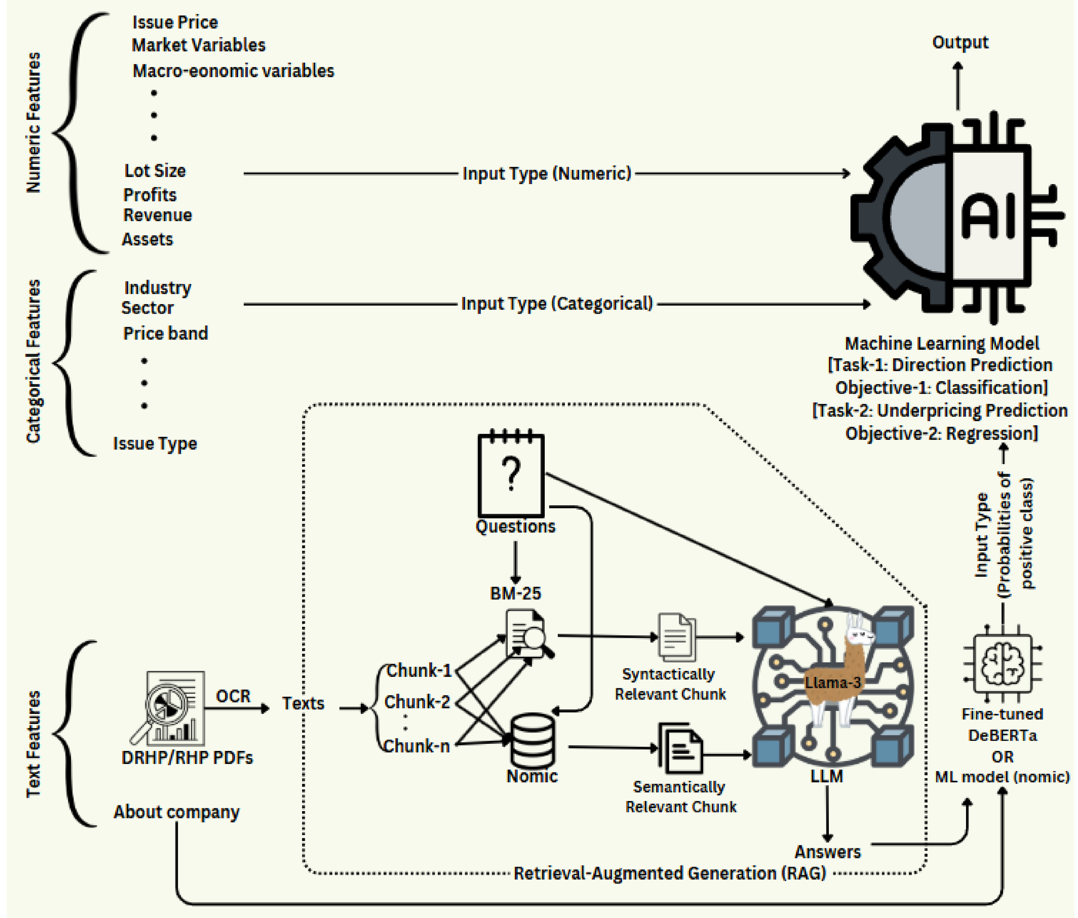
FIRE-2023

Task-13: Extract ESG Issues, Assess Sustainability, and Detect exaggerated numerals from FTs in Hindi, Bengali, & Telugu

LREC-COLING 2024

Task-14: Predicting direction and under-pricing with respect to Open, High, Close prices of Indian IPOs

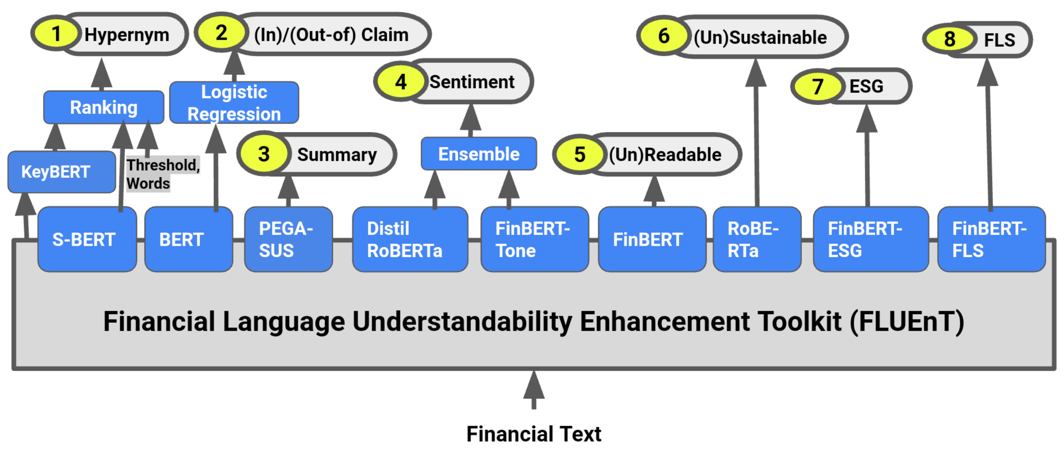
arXiv:2412.16174



Tools for FinNLP

CODS-COMAD 2023

Task-15: Financial Language Understandability Enhancement Toolkit (FLUEnT), ESG Issue Detector (EID), Financial Claim Analysis Tool (FinCAT), etc.



Approaches and results for different tasks

AU-ROC = Area under the ROC curve, Acc. = Accuracy, MPP = Maximum Possible Profit, ML = Maximum Loss, MAPE = Mean Absolute Percentage Error, NA = Not Applicable, SOTA = State of the Art, LLM = Large Language Model, PLM = Pre-trained Language Model, Trans-Prp = Translate Paraphrase, IT = Impact Type, ID = Impact Duration, McL = Machine Learning, Num = Numeric Features, Cat = Categorical features, Txt = Text Features

CIKM-2024

Task #	Metric	Approach Summary	SOTA	Performance	New Data	Language	New Tool
1	AU-ROC	FinBERT finetune	Yes	0.993	Yes	English	Yes
2	F1	RoBERTa + Claude (LLM)	Yes	0.731	Yes	English	No
3	Acc.	SBERT finetune	Yes	0.967	No	English	No
4	F1	SEC-BERT + Neural Network	No	0.736	No	English	No
5	Acc.	RoBERTa finetune	No	0.932	No	English	No
6	F1	SEC-BERT finetune	No	0.715	No	English	Yes
7	F1	FinBERT finetune	No	0.929 (IT)	No	English	No
7	F1	Trans-Prp + FinBERT finetune	No	0.756 (IT)	No	French	No
7	F1	Trans-Prp + FinBERT finetune	Yes	0.679 (IT)	No	Japanese	No
7	F1	Trans-Prp + FinBERT finetune	Yes	0.677 (IT)	No	Chinese	No
7	F1	Trans-Prp + PLM finetune	No	0.5882 (ID)	No	English	No
7	F1	Trans-Prp + PLM finetune	Yes	0.5616 (ID)	No	French	No
8	F1	Ensemble (FinBERT, BERT + Logistic Regression)	No	0.948	No	English	Yes
9	MPP, ML	SBERT Chinese + Classifier, FinBERT	No	0.575 (MPP), 0.598 (ML)	No	Chinese	No
10	MAPE	Gated Recurrent Unit	Yes	0.382	Yes	English	Yes
11	F1	Cross Encoder (FinBERT Finetuned)	No	0.789	No	English	No
11	F1	Translate + Cross Encoder (SEC-BERT)	No	0.641	No	Chinese	No
12	F1	MBERT, Cross Encoder (MBERT)	No	0.721 (1st task), 0.755 (2nd Task)	Yes	Bengali	Yes
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.680 (1st task), 0.950 (2nd task), 0.590 (3rd task)	Yes	Hindi	No
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.650 (1st task), 0.920 (2nd task), 0.550 (3rd task)	Yes	Bengali	No
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.680 (1st task), 0.920 (2nd task), 0.580 (3rd task)	Yes	Telugu	No
14	F1	McL (Num, Cat, Txt) (Classification)	Yes	0.947 (Open-MB), 0.935 (High-MB), 0.931 (Close-MB)	Yes	English	No
14	MAE	McL (Num, Cat, Txt) (Regression)	Yes	0.167 (Open-MB), 0.193 (High-MB), 0.194 (Close-MB)	Yes	English	No
14	F1	McL (Num, Cat, Txt) (Classification)	Yes	0.893 (Open-SME), 0.942 (High-SME), 0.911 (Close-SME)	Yes	English	No
14	MAE	McL (Num, Cat, Txt) (Regression)	Yes	0.239 (Open-SME), 0.263 (High-SME), 0.256 (Close-SME)	Yes	English	No
15	NA	Gradio (frontend)	NA	NA	NA	Various	Yes

Venues: TheWebConf (WWW), CIKM, LREC-COLING, etc.
Travel Grants: CODS-COMAD, CIKM, IndoML, PIC, ARCS

CODS COMAD YRS-2023 (Honourable Mention)