LIPI at the NTCIR-16 FinNum-3 Task: Ensembling transformer based models to detect in-claim numerals in Financial Conversations

Sohom Ghosh [1,2] Sudip Kumar Naskar [2]

[1]*Fidelity Investments, India* [2]Jadavpur University, India

Introduction

Investors tend to make data - decisions by consuming financial content available online. Numerals present in these contents may be simply claims by executives and not facts. Our motivation is to help investors to identify out-of-claim numerals so that they do not get allured by in-claim numerals. For example, in the Figure 1, 2021 is out-of-claim whereas 80% is in-claim. We present a system to classify numerals occurring in financial texts as in-claim or out-of-claim.

Figure 1. Numeral classification in Financial Texts

Exploratory Data Analysis

Our training, validation and test set comprised 8337, 2383 and 1191 instances respectively. We present the word clouds of each class for training and validation sets in Figure 2. Moreover, we also studied how the number of sentences and tokens varied across different sources. These are presented in Figures 3 and 4. We observe that the number of sentences and tokens for the validation set is lesser than that of the training and test set for each of the categories.

We followed the instructions provided the organizers of FinNum-3 and used Micro and Macro F1 scores for evaluation.

Figure 2. Word clouds of Training and Validation sets for each class

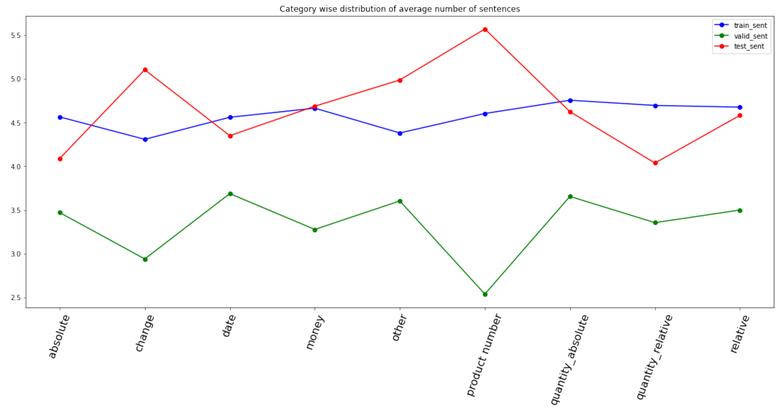

Figure 3. Distribution of number of sentences

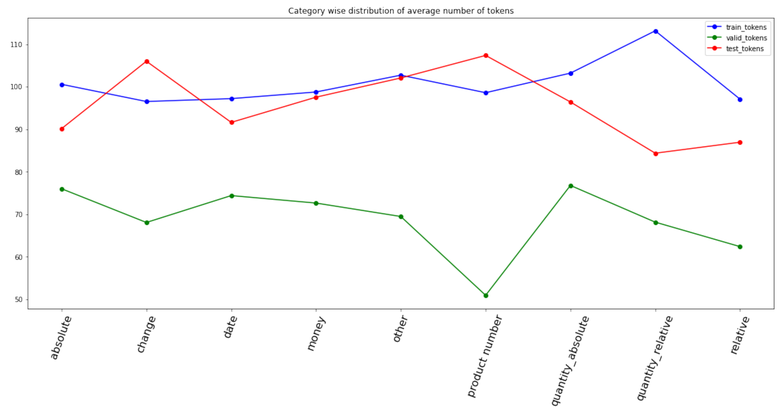

Figure 4. Distribution of number of tokens

Methodology

Our final system is an ensemble of three systems. It uses majority voting to decide the final output.

For the first model M1, after lots of experimentation, we decided to use words within a range of 8 before and after the target numeral as the context. We fine-tuned a FinBERT model for 15 epochs considering a maximum of 64 tokens for the task of classification.

In the second model M2, we reduced the range from 8 to 6 and the number of maximum tokens from 64 to 16. Everything else was the same.

For the third model M3, we extracted 6 words proceeding and succeeding the target numeral. We referred to this as the context window. Given this context window, we calculated the average of the BERT embeddings of the constituent tokens present in the target numeral. This means that the contextual embeddings of the target numeral having 768 dimensions were used as features. Additionally, we used some engineered features (like one-hot vectors of the categories extracted using Microsoft Recognizers for Text, the number of digits before and after the decimal, parts of speech of words preceding and succeeding the target numeral and so on.) to train the Logistic Regression model.

Results and Discussions

We made three submissions, the first one is the output of model M1 as it is. The second one is similar to the third one. The only difference is we fine-tuned the threshold for classification so that the model performs the best on the validation set without compromising much on the training set. Our third submission is the output of the ensemble model M3. We present the results in Table 1.

Table 1. Results

	Validation		Test	
submission	**Micro-F1**	**Macro-F1**	**Micro-F1**	**Macro-F1**
LIPI_1	94.45%	85.84%	95.09%	82.82%
LIPI_2	**95.80%**	**88.15%**	95.17%	81.33%
LIPI_3	94.79%	86.71%	**95.59%**	**84.73%**

Table 2. Ablation Study. CW = Context Window, EF = Engineered Features

Model	**Macro-F1**	**Micro-F1**
M3 (only)	0.8318	0.6646
M3 (-EF)	0.8238	0.7990
FinBERT classifier (CW=5)	0.8318	0.7250
FinBERT classifier (CW=6)	0.8439	0.6603
FinBERT classifier (CW=7)	0.8381	0.8244
FinBERT classifier (CW=8) (`LIPI_1`)	0.8585	0.6345
FinBERT classifier (CW=9)	0.8247	0.8262
Ensemble (`LIPI_3`)	**0.8671**	**0.9479**

We studied the effect of changing the range within which words surrounding the target numeral were considered. Furthermore, we experimented by removing the hand crafted features from M3. This is presented in Table 2.

References

[1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

[2] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Numclaim: Investor's fine-grained claim detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1973–1976, New York, NY, USA, 2020. Association for Computing Machinery.

[3] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-16 finnum-3 task : Investor's and manager's fine-grained claim detection, 2022.

[4] Microsoft Corporation. Microsoft recognizers text. `https://github.com/microsoft/Recognizers-Text`, 2018. [Online; accessed Dec-2021].

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Sohom Ghosh and Sudip Kumar Naskar. Fincat: Financial numeral claim analysis tool. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*, New York, NY, USA, April 2022. Association for Computing Machinery.

Summary

The third edition of FinNum shared task, being held with NTCIR-16 presented the challenge of classifying numerals present in financial texts into in-claim or out-of-claim classes. It consisted of two claim detection sub-tasks on i) professional analysts' reports written in Chinese and ii) earning conference calls transcribed in English. In this paper, we describe the approach our team (LIPI) followed while participating in the English subtask of FinNum-3. This approach consists of ensembling transformer based models with a Logistic Regression model trained using BERT embeddings and hand-crafted features. It out-performed the existing baseline and achieved a macro F1 score of 84.73% and micro F1 score of 95.59% on the test set.