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

Sohom Ghosh
Fidelity Investments
Bengaluru, Karnataka, India
Jadavpur University
Kolkata, West Bengal, India
sohom1ghosh@gmail.com

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

ABSTRACT

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.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

claim detection, financial texts, call transcripts

TEAM NAME

LIPI

SUBTASKS

FinNum-3: Investor's and Manager's Fine-grained Claim Detection (English)

1 INTRODUCTION

Detecting in-claim and out-of-claim numerals from financial documents are extremely essential. This aids investors to make proper decisions. This sub-task [3] provide us with formal financial documents. These documents include transcripts from earnings conference calls (in English) and reports written by professional analysts (in Chinese [2]). We aimed to develop a system that takes numerals present in call transcripts as inputs and to classify them as in-claim or out-of-claim. We depict this in Figure 1. The problem statement has been narrated formally in the paper [3]. In-calim numerals are represented as class 0 and out-of-claim numerals are represented as class 1. 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 4. Moreover, we also studied how the number of sentences and tokens varied across different sources. These are presented in Figures 2 and 3. 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.

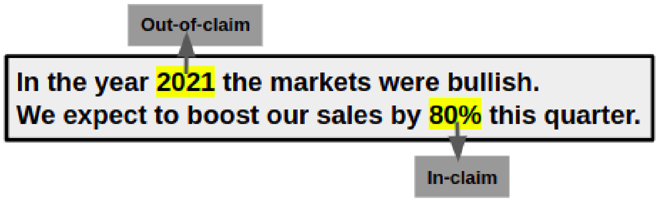

Out-of-claim
In the year 2021 the markets were bullish.
We expect to boost our sales by 80% this quarter.
In-claim

Figure 1: Numeral classification in Financial Texts

2 SYSTEM DESCRIPTION OF THE SUBMISSIONS

The number of submissions for each team was restricted to 3. In this section, we narrate each of our submissions.

2.1 Submission-1 (LIPI_1)

After analysing the training data, we realized that most of them have more than one numeral. Thus, we needed to consider words within a certain range of the target numeral. We experimented with several ranges and finally decided to use words within a range of 8 before and after the target numeral. Using the standard procedure, we fine-tuned a FinBERT [1] model for 15 epochs considering a maximum of 64 tokens for the task of classification.

2.2 Submission-2 (LIPI_2)

This submission was almost identical to the next one (i.e. submission 3). The only difference was that 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.

2.3 Submission-3 (LIPI_3)

This was our best performing model. It was an ensemble model consisting of three individual models. The first model was the one described as submission-1 (LIPI_1). In the second model, we reduced the range from 8 to 6 and the number of maximum tokens from

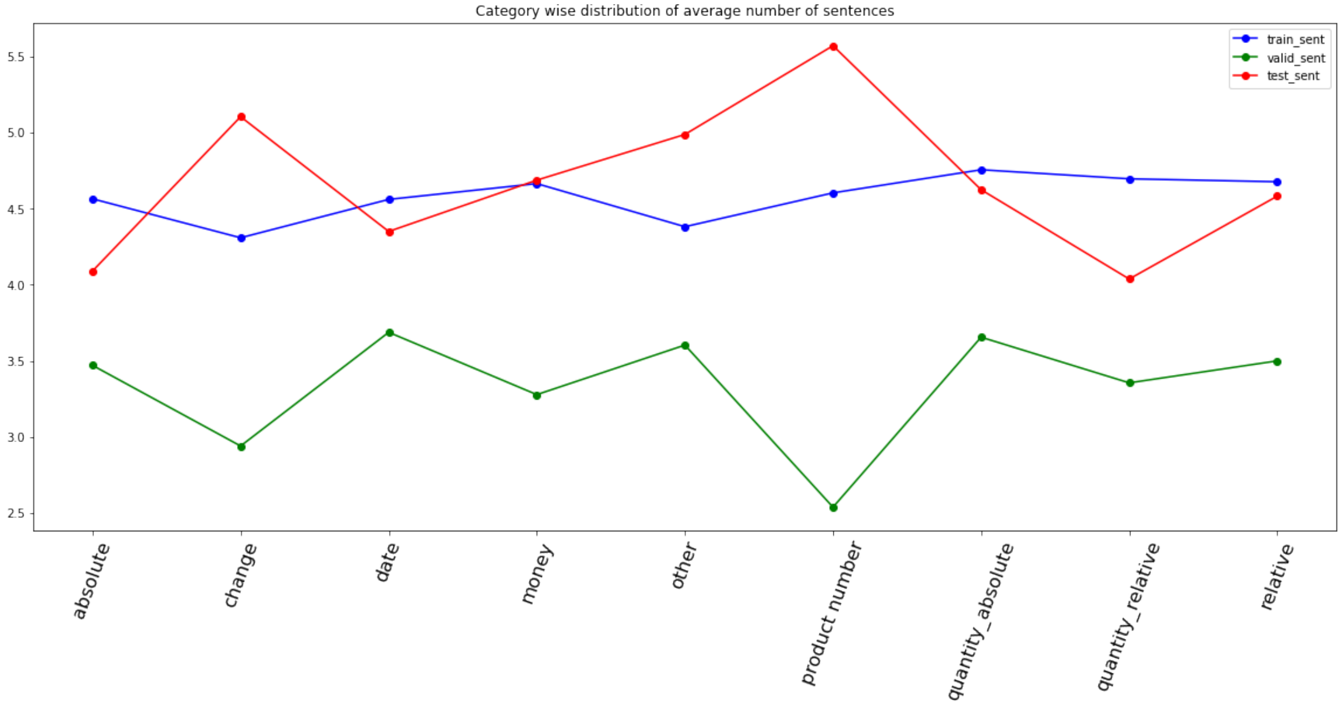


Figure 2: Distribution of number of sentences

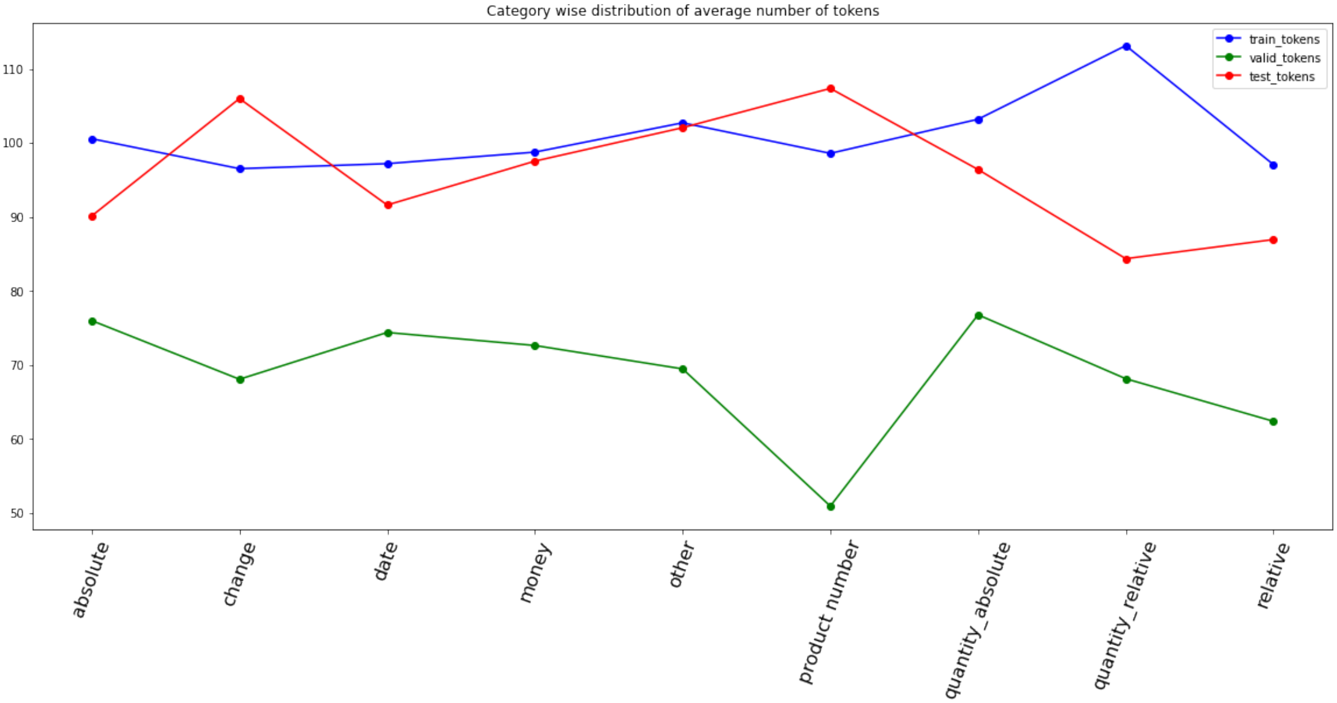
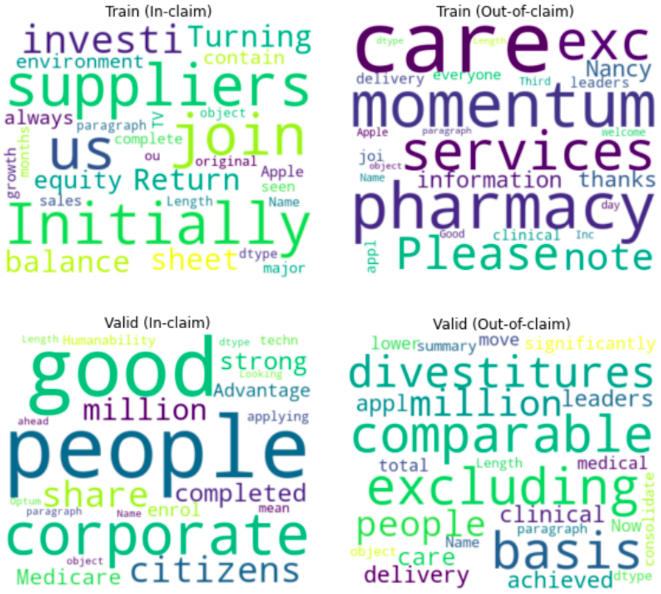


Figure 3: Distribution of number of tokens


Train (In-claim)

Train (Out-of-claim)

Valid (In-claim)

Valid (Out-of-claim)

Figure 4: Word clouds of Training and Validation sets for each class

64 to 16. Everything else was the same. In the third model ***M3***, we used 768-dimensional context-based BERT [5] embedding of the target numeral for a context range of 6 words occurring before and after it. In addition to this, we used several hand-crafted features. This includes

- one-hot vectors of the categories extracted using Microsoft Recognizers for Text [4]
- number of digits before and after the decimal
- parts of speech of words preceding and succeeding the target numeral.

We trained a Logistic Regression model using these features. We used 0.5 as the threshold for classification. Lastly, we decided on the final class of the target numeral using majority voting.

We present the results for each of our submissions in Table 1. The performance numbers on the test set have been provided by the organizing team. On comparing the performance of `LIPI_3` with that of other participants we observe that it ranked 9th and 10th out of 16 submissions in terms of micro and macro F1 scores respectively.

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%**

3 ABLATION STUDY

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***. Table 2 presents the findings.

4 CONCLUSIONS

In this paper, we discussed the approach our team LIPI followed while participating in the NTCIR FinNum-3 (English) shared task. In terms of Macro F1 score, our best model improved the baseline by 47.72%. We have also developed a tool named “FiNCAT” [6] which could detect claims in financial texts in real-time. However, four other teams performed better than us.

In future, we would like to analyse the false positives and true negative instances. Furthermore, we want to benchmark our methodology with that of the other participants. This will help us to get ideas for further improving the performance of our model. Another interesting direction would be to investigate how significant is the performance gain of the winning team as compared to the others.

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**

5 ACKNOWLEDGEMENTS

We would like to thank the organizers for providing us with the labelled datasets and resolving our queries from time to time.

REFERENCES

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

[2] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NumClaim: Investor’s Fine-Grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM ’20)*. Association for Computing Machinery, New York, NY, USA, 1973–1976. https://doi.org/10.1145/3340531.3412100

[3] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Overview of the NTCIR-16 FinNum-3 Task : Investor’s and Manager’s Fine-grained Claim Detection. forthcoming

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

[5] 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

[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