# Overview of the NTCIR-15 FinNum-2 Task: Numeral Attachment in Financial Tweets

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Department of Computer Science, National Chengchi University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@nccu.edu.tw

Hiroya Takamura
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan
takamura.hiroya@aist.go.jp

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

## ABSTRACT

In FinNum task series, we focus on understanding the numeral-related information in the financial social media data. In FinNum-2, we introduce a new task, named numeral attachment, to identify the relation between the mentioned stock and the numerals in a financial tweet and propose a NumAttach 2.0 dataset with 10,340 expert-annotated instances. In this paper, we give an overview of FinNum-2 shared task and analyze the results of 17 submissions from 7 teams. The statistics of NumAttach 2.0 and the comparison of participants' results with that of baseline models are provided.

## CCS CONCEPTS

• **Information systems** → **Information extraction**.

## KEYWORDS

Numeral attachment, financial social media, numeral corpus

## 1 INTRODUCTION

Recently, the numerals in the documents attract lots of attention. Some researches test the numeracy of neural network models in the narratives of different domains [6, 17, 18, 20]. Some works focus on mining the numeral information in the textual data [5, 13]. These works show that the language model of numerals may be different from those designed for words and indicate the importance of dealing with numeral information. In FinNum shared task series, we pay attention to the numerals in financial social media data.

Last year, we introduce a numeral understanding task [5] and publish FinNum 2.0 dataset [7]. With FinNum 2.0 dataset, the meaning of the given numeral could be disambiguated by the proposed methods of FinNum-1 participants. However, when we try to apply the results of FinNum-1 to real-world applications [1, 8], the other problem, called numeral attachment [2], is raised. That is, there may exist more than one numeral or more than one cashtag (special tag for the mentioned stock) in a tweet on financial social media platforms. In the multi-numeral or multi-cashtag cases, we need to identify whether the given numeral is related to the given cashtag. Figure 1 provides an example with one cashtag and two numerals.

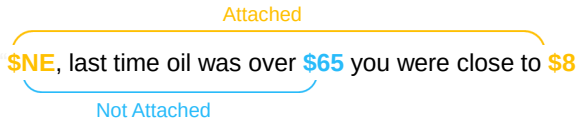


**Figure 1: Example of numeral attachment.**

**Table 1: Statistics of NumAttach.**

| | Attached | Not Attached | Sum |
|---|---|---|---|
| Train | 5,827 | 1,360 | 7,187 |
| Development | 850 | 194 | 1,044 |
| Test | 1,721 | 388 | 2,109 |
| Sum | 8,398 | 1,942 | 10,340 |

The numeral "8" is the price of $NE (Noble Corporation plc). The numeral "65" is the price of oil but not related to the cashtag $NE. In this case, if we do not deal with the attachment issue, both "8" and "65" will be considered as the price of $NE. That may cause the wrong estimation of downstream analysis. In order to solve this problem, in FinNum-2, we propose a new dataset, NumAttach, for solving the numeral attachment problem on financial social media platforms.

The rest of this paper is organized as follows. In Section 2, we provide the statistics of NumAttach dataset. In Section 3, we highlight the characteristics of the models proposed by FinNum-2 participants. In Section 4, we provide the experimental results in official runs. In Section 5, we separate the test data into different classes and analyze the results of different models on these classes. In Section 6, we conclude the findings in FinNum-2 shared task.

## 2 DATASET

Table 1 shows the statistics of the proposed dataset. There are a total of 10,340 instances in the NumAttach dataset. We use 70% of data as training set, 10% of data as development set, and 20% of data

**Table 2: Distribution of different kinds of instances.**

| | Single-cashtag | Multi-cashtag |
|---|---|---|
| Single-numeral | 1,282 (12.40%) | 1,427 (13.80%) |
| Multi-numeral | 3,347 (32.37%) | 4,284 (41.43%) |

**Table 3: The labels (Attached/Not Attached) of different kinds of instances.**

| | Single-cashtag | Multi-cashtag |
|---|---|---|
| Single-numeral | 1,204/78 | 1,017/410 |
| Multi-numeral | 3,071/276 | 3,106/1,178 |

**Table 4: Distribution of cashtag and numeral.**

| # of | Cashtag | Numeral |
|---|---|---|
| 1 | 44.77% | 26.20% |
| 2 | 26.56% | 28.33% |
| 3 | 10.64% | 17.87% |
| 4 | 5.90% | 10.11% |
| 5 | 3.43% | 6.03% |
| 6 | 2.40% | 3.43% |
| 7 | 1.54% | 3.37% |
| 8 | 0.98% | 1.05% |
| 9 | 0.88% | 2.13% |
| > 10 | 2.91% | 1.48% |

as test set. Table 2 shows the distribution of different kinds of instances, including single-/multi-cashtag and single-/multi-numeral. Only 12.40% of data is single-cashtag and single-numeral cases. We further show the statistics of the labels in different kinds of instances in Table 3. Most of the cashtags and the numerals in the single-cashtag/single-numeral cases are related, i.e., attached. However, in the multi-cashtag or multi-numeral cases, there are many "not attached" cases. The above-mentioned phenomena evidence the importance of the proposed task.

In Table 4, we show the distribution of cashtag and numeral. Over 87% and 82% of financial tweets contain less than 4 cashtags/numerals, respectively. Most instances with more than 4 cashtags are used for listing the stocks with similar characteristics. Here is an example for this kind of financial tweets: "Nov 29th Watchlist: Swing entries for $EYEG $OPTT Love $HUSA Tomorrow. Continuation $NURO $BIOC $MNKD $CLNT Option $DPZ $KEM $GE #TradeSmart". In this kind of cases, most numerals are related to date or time information and are not attached to any cashtag. That shows the number of cashtags in a tweet may be a feature for solving the numeral attachment problem.

## 3 METHODS IN OFFICIAL RUNS

In FinNum-2, most participants use transformer-based [19] models, especially, BERT [10] and RoBERTa [15]. Chen and Liu (MIG) [9] explore some features and different settings such as class weight with BERT. Liang et al. (TMUNLP) [14] probe the results of adding different architectures before BERT encoder , and experiment on adding the dependency features. Moreno et al. (TLR) [16] and Jiang et al. (CYUT) [12] experiment on both BERT and RoBERTa. Additionally, Moreno et al. [16] propose an ensemble approach (TLR-3) by considering the outputs of both BERT and RoBERTa simultaneously, which performs the best in the test set. Xia et al. (WUST) [21] provide the results of using the SVM model.

**Table 5: Experimental results. (%) Dev. denotes the results of development set. Dep. denotes the dependency features. CW denotes the class-weight setting.**

| Team | Method | Dev. | Test |
|---|---|---|---|
| Baseline - 1 | Majority | 44.88 | 44.93 |
| CYUT-1 | - | 48.64 | 48.02 |
| WUST | SVM | 82.91 | 54.43 |
| BTBCH-1 | - | 100.00 | 57.19 |
| BTBCH-2 | - | 99.68 | 58.00 |
| TMUNLP-3 | BERT-CNN + Dep. | 87.34 | 58.40 |
| TMUNLP-2 | BERT-BiLSTM + Dep. | 85.17 | 59.77 |
| IIITH-1 | - | 96.16 | 62.81 |
| Baseline - 2 [2] | Caps-m | 79.27 | 63.37 |
| IIITH-3 | - | 93.99 | 64.16 |
| TMUNLP-1 | BERT-BiLSTM | 87.02 | 64.76 |
| MIG-1 | BERT-BiLSTM + CW | 84.46 | 68.27 |
| MIG-3 | BERT + CW | 90.69 | 68.37 |
| TLR-2 | RoBERTa | 87.81 | 68.64 |
| MIG-2 | BERT-BiLSTM + CW | 85.77 | 68.72 |
| IIITH-2 | - | 96.23 | 71.11 |
| TLR-1 | BERT | 88.26 | 71.41 |
| CYUT-2 | RoBERTa | 95.99 | 71.90 |
| TLR-3 | Ensemble | 88.87 | **73.95** |

## 4 EXPERIMENTAL RESULTS

In this paper, we use the macro-F1 score[1] to evaluate the experimental results. We report the results of two baseline methods, including the results of the majority and the capsule-based model (Caps-m) in our previous work [2]. Table 5 shows the results of baseline methods and the results of participants' models. We find that the vanilla BERT and RoBERTa achieve good performances. The difference between the performances of different teams using the same architecture may be caused by the preprocessing procedures and the hyperparameter settings.

In Table 6, we show the correct rate of the test set. The correct rates of 79% instances are larger than 80%, and only 10% instances have the correct rate lower than 20%. Taking a closer look at the instances with a lower correct rate, in Table 7, we find that all of these instances are labeled as "Not Attached". Based on the statistics in Table 1, the ratios of "attached" labels and "not attached" labels are 82.22% and 18.78%, respectively. The highly unbalanced between labels may be the reason that caused the lower correct rate on the "not attached" instances. We further show the averaged correct rate on different kinds of instances in Table 8. We find that the multi-numeral cases may be easier than single-numeral cases, and the multi-cashtag cases are harder than single-cashtag cases.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

**Table 6: Distribution of correct rate.**

| Correct Rate | # of Instance | % |
|---|---|---|
| 0.00% | 107 | 5.07% |
| 5.56% | 46 | 2.18% |
| 11.11% | 43 | 2.04% |
| 16.67% | 32 | 1.52% |
| 22.22% | 15 | 0.71% |
| 27.78% | 10 | 0.47% |
| 33.33% | 17 | 0.81% |
| 38.89% | 18 | 0.85% |
| 44.44% | 8 | 0.38% |
| 50.00% | 23 | 1.09% |
| 55.56% | 13 | 0.62% |
| 61.11% | 17 | 0.81% |
| 66.67% | 24 | 1.14% |
| 72.22% | 38 | 1.80% |
| 77.78% | 33 | 1.56% |
| 83.33% | 110 | 5.22% |
| 88.89% | 216 | 10.24% |
| 94.44% | 1339 | 63.49% |

**Table 7: The labels (Attached/Not Attached) of instances with correct rate lower than 20%.**

| | Single-cashtag | Multi-cashtag |
|---|---|---|
| Single-numeral | 0/64 | 0/10 |
| Multi-numeral | 0/152 | 0/2 |

**Table 8: Averaged correct rate on different kinds of instances.**

| | Single-cashtag | Multi-cashtag |
|---|---|---|
| Single-numeral | 79.78% | 55.17% |
| Multi-numeral | 81.68% | 67.46% |

# 5 DISCUSSION

## 5.1 Item Difficulty Index

Leveraging to the results of the participants, we calculate the difficulty and discrimination by comparing the results of top-5 models and the last-5 models. With these scores, future works can focus on dealing with hard instances. The difficulty (DIF) of a instance and the discrimination (DIS) of a instance are calculated based on the correct rate (CR) as follows:

$$\text{DIF} = \frac{CR_{top-k} + CR_{last-k}}{2} \tag{1}$$

$$\text{DIS} = CR_{top-k} - CR_{last-k}\,, \tag{2}$$

where $k$ is the number of models. In this paper, $k$ is 5. where $k$ is the number of models. In this paper, we set $k$ as 5. Table 9 and Table 10 show the statistics of DIF and DIS. Based on Ebel [11], if the DIS of an instance is higher than 0.4, we say that this instance is a good instance for discriminating the models.

**Table 9: Difficulty of test instances. The lower difficulty score means the instance is harder.**

| Difficulty | # of Instance | % |
|---|---|---|
| 0.0 | 140 | 6.64% |
| 0.1 | 46 | 2.18% |
| 0.2 | 47 | 2.23% |
| 0.3 | 35 | 1.66% |
| 0.4 | 37 | 1.75% |
| 0.5 | 40 | 1.90% |
| 0.6 | 45 | 2.13% |
| 0.7 | 115 | 5.45% |
| 0.8 | 143 | 6.78% |
| 0.9 | 1461 | 69.27% |

**Table 10: Discrimination of test instances.**

| Discrimination | # of Instance | % |
|---|---|---|
| -0.6 | 3 | 0.14% |
| -0.4 | 23 | 1.09% |
| -0.2 | 47 | 2.23% |
| 0.0 | 191 | 9.06% |
| 0.2 | 1536 | 72.83% |
| 0.4 | 158 | 7.49% |
| 0.6 | 94 | 4.46% |
| 0.8 | 40 | 1.90% |
| 1.0 | 17 | 0.81% |

## 5.2 Error Analysis

As we shown in Table 7, all instances with lower correct rate is "Not Attached" cases. Table 11 shows some instances with lower correct rate. As shown in the first instance, informal writing style is still one of the main challenges when analyzing the social media data. In the second instance, if models refer to the history price of $JWN ($12-24), it may be earlier to find that $250 may not be related to $JWN. The third instance shows that the target numeral is attached to other named entity, which is not represented by a cashtag, in the tweet. It points out one of the future research directions — general numeral attachment task. That is, in FinNum-2, we take advantage of using cashtag to detect the target named entity. In general numeral attachment task, researchers should try to detect the attached named entity given a target numeral. The forth instance shows that the temporal information (Q4) may not be directly related to $BOX, but it points out the execute date of $BOX's deals. This may be an ambiguous case.

## 5.3 Future Direction

Understanding the category (FinNum-1) and the attached entity (FinNum-2) of a numeral are fundamental tasks for fine-grained numeral understanding. How to extract the numeral containing investor's subjective opinion and find the rationale provided to support investor's opinion are still an open issue for investor's opinion mining. Based on our observations [3], most social media users on Twitter or StockTwits do not provide the rationales to support their opinions, which makes researchers hard to explore

**Table 11: Cases with low correct rate.**

| Example | Ans | Cause |
|---|---|---|
| **$IMUC** head banging against the ceiling. Matter of time b**4** she breaks wide open! | Not Attached | Informal writing style |
| **$JWN** millennials will not spend **$250** for jeans at Nordstrom. | Not Attached | Need price data for inference |
| **$GLD** VIX under **10**? Markets will sell in Oct, Bonds stink. | Not Attached | Attached to not-cashtag entity |
| Already in Q**4**, **$BOX** is set to execute a record number of big deals | Not Attached | Ambiguous |

more fine-grained tasks on these data. Thus, we think that more fine-grained tasks can be probed (1) with the social media data with longer narratives such as blog articles or (2) with the formal documents such as professional analysis report. Although the pilot dataset, NumClaim [4], is proposed for extracting the fine-grained claim from professional analysis reports, we think that the premise extraction and claim-premise relation linking are still worth exploring. Besides, how to sort out the trustworthy or high-quality opinions is also a possible extension of the FinNum shared tasks. We also believe that the proposed numeral understanding tasks and the findings of participants can be extended to understand the numerals in other domain such as clinical documents.

## 6 CONCLUSION

In this paper, we provide an overview of FinNum-2 shared task in NTCIR-15. The expert-annotated dataset, NumAttach, is published. Several solutions for the numeral attachment in financial social media data are proposed. Due to the unbalanced label distribution and the informal writing style, some instances still cannot be solved. With our analysis and the experiences of FinNum-2 participants, future works can focus on the harder instances we found and further improve the performance of numeral attachment.

## ACKNOWLEDGMENTS

We greatly appreciate the efforts of all the participants in the FinNum-2 shared task at NTCIR-15. This shared task was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 109-2634-F-002-040, and MOST 109-2634-F-002-034, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

## REFERENCES

[1] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Crowd View: Converting Investors' Opinions into Indicators.. In *IJCAI*. 6500–6502.
[2] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1161–1164.
[3] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and Perspectives from 10,000 Annotated Financial Social Media Data. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 6106–6110.
[4] 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*. 1973–1976.
[5] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 136–143.
[6] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6307–6313. https://doi.org/10.18653/v1/P19-1635
[7] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*. 19–27.
[8] Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. 2019. Crowdpt: Summarizing crowd opinions as professional analyst. In *The World Wide Web Conference*. 3498–3502.
[9] Yu-Yu Chen and Chao-Lin Liu. 2020. MIG at the NTCIR-15 FinNum-2 Task: Use the transfer learning and feature engineering for numeral attachment task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.
[10] 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
[11] Robert L Ebel and David A Frisbie. 1972. *Essentials of educational measurement*. Prentice-Hall Englewood Cliffs, NJ.
[12] Mike Tian-Jian Jiang, Yi-Kun Chen, and Shih-Hung Wu. 2020. CYUT at the NTCIR-15 FinNum-2 Task: Tokenization and Fine-tuning Techniques for Numeral Attachment in Financial Tweets. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.
[13] Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky, and Percy Liang. 2018. Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 82–92. https://doi.org/10.18653/v1/D18-1008
[14] Yu-Chi Liang, Yi-Hsuan Huang, Yu-Ya Cheng, and Yung-Chun Chang. 2020. TMUNLP at the NTCIR-15 FinNum-2. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.
[15] 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. *arXiv preprint arXiv:1907.11692* (2019).
[16] Jose G. Moreno, Emanuela Boros, and Antoine Doucet. 2020. TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.
[17] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring Numeracy in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3374–3380. https://doi.org/10.18653/v1/P19-1329
[18] Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2104–2115. https://doi.org/10.18653/v1/P18-1196
[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[20] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. 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, 5307–5315. https://doi.org/10.18653/v1/D19-1534
[21] Xinxin Xia, Wei Wang, and Maofu Liu. 2020. WUST at NTCIR-15 FinNum-2 Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.