Passau21 at the NTCIR-16 FinNum-3 Task: Prediction Of Numerical Claims in the Earnings Calls with Transfer Learning

Alaa Alhamzeh
Universität Passau
Germany
alaa-alhamzeh@uni-passau.de

M. Kürsad Lacin
Universität Passau
Germany
lacin01@ads.uni-passau.de

Előd Egyed-Zsigmond
INSA de Lyon
France
Elod.Egyed-zsigmond@insa-lyon.fr

ABSTRACT

The FinNum Task series aims at better understanding of numeral information in financial narratives. The goal of FinNum-3; on the English data part; is to have a fine-grained manager's claim detection in the Earning Conference Calls (ECCs) with the help of Natural Language Processing (NLP). To succeed in the best performance for predicting in-claim and out-of-claim numerals, we propose the BERT (Bidirectional Encoder Representations from Transformers) base model , which is pre-trained on a large corpus of English data. The results of our model are 86.48% of macro-F1 score in the validation split and 87.12% of macro-F1 score in the test data.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
• **Information systems** → **Information retrieval**.

KEYWORDS

FinNum-3, Earnings Conference Calls, Financial Claim Detection

TEAM NAME

PASSAU21

SUBTASKS

Manager's claim detection (English)

1 INTRODUCTION

A better automatic understanding of financial data is always challenging given its particular genre and structure of text. More specifically, the frequent usage of numerals in financial narratives versus other domains makes it an urgent task to efficiently understand the numerals as a core-stone for analysing the financial documents. Recently, researchers with different backgrounds such as computer science, economics, and finance proposed different approaches to improve financial application and data analysis using the recent advancement of technology. This is known as FinTech (Financial Technology). In this regard, FinNum task series tackled this challenge and aimed at interpreting numerical information sufficiently in different data sources and different languages. In FinNum-1 [1], fine-grained numeral understanding in social media is analyzed by 9 participants. The FinNum-2 [2] task investigated the numeral attachment in financial tweets. The English part of the third edition of FinNum series is designed to study the numeral data in Earnings Conference Calls (ECCs) using Argument Mining (AM) from Natural Language Processing (NLP) field.

In fact, the earning conference calls represents one of the most important data sources for financial applications like: stock movement prediction, volatility forecasting, estimation of analysts recommendations. However, most of the literature analyse them using basic linguistic features and sentiment analysis. While in the task proposal, the goal is to focus weather a mentioned number is part of argumentation process or not.

Therefore, the task goal is, briefly, to examine a given number in a sentence of managers' speeches whether it is part of a claim or not, i.e., as a binary classification task. Even though the premises and their relations with claims are incredibly crucial in argument mining, the data considers only the claims. Additionally, the data also includes the category information of the target numeral, for example, date, monetary, and others, which itself was the task of FinNum-1. The task organizers allow participants to decide whether to use this information to design joint learning models. However, after some data analysis, we chose not to use the category. Therefore, we are going to explain in this paper, our proposed model to solve the claim detection task as a flat problem.

Chen et al.[3] claimed that the earnings conference calls have a significant role in announcing some of company private information and earnings reports to the public and hence changing analysts' recommendations and crowd expectations. Researchers investigated the stock volatility and price prediction based on the earnings calls data in the last decade. However, at the end of each earning call, there's the question and answering session (QA) where the company representatives answer the questions of professional analysts and business journalists. Consequently, this QA parts of the ECCs are not well structured and pre-scripted as in the presentation part, but they are still full of arguments that the company representatives use to defend their views and convince the other party to believe in them. Hence, this section of ECCs is more promising for argument mining analysis.

On the contrary, Keith et al.[4] proved that the usage of only QA parts as not encouraging as whole ECC for a sentiment analysis based method. More recently, researchers has also considered the vocal features of earnings calls as well as verbal features [5–8]. The combination has better performance on the volatility prediction than only the transcript investigation.

Numerals have a more notable role in the financial narrative than the other domains [3]. Thus, understanding and analyzing the numerals in the ECCs will give us a satisfactory interpretation of

how well the company managers handle the ECCs. However, not all numerals are essential in consideration of expectation. Instead, we are interested in numbers that are part of arguing and claiming process. The detection of those in-claim numerals is the primary objective of FinNum-3, and in this paper we present our proposed solution to this end.

This paper is organized as follows: Section 2 represents the a conceptual background and related work. In Section 3, we describe the task more in details. The dataset preparation is discussed in Section 4 as well as our proposed solution using transfer learning model. We then present the evaluation results and analyse model performance in Section 5. Finally, conclusion and the future work are exhibited in Section 6.

2 BACKGROUND AND RELATED WORK

The FinNum task series is interested in numerals in finance, as we can understand from the abbreviation. Furthermore, in FinNum-3, argument mining with claim detection in the earnings calls is the main task. Therefore, we present in the following, an overview about each of argument mining and ECC and we state some recent studies considering them.

2.1 Argument mining

Argumentation is the logical reasoning people use to reach out a conclusion. The simplest form of an argument is a combination of one premise supporting a final claim or a conclusion [9]. Argument mining, which tackles the automatic detection of arguments in an input text, has gained a lot of research interest in the last decade given the wide spectrum of applications it implies (e.g., assisted writing, search engines, legal text, decision making, etc.). However, the main challenge of argument mining is that the structure of arguments is very domain dependant. Therefore, many solutions have been addressed to build cross-domain models that can still be able to correctly classify arguments over multiple datasets, from different data sources. One recent example is the work introduced by Alhamzeh et al. [10] where they adopt the concept of ensemble learning to combine the classification outputs of a classical machine learning classifier: SVM [11], and a fine-tuned transfer learning model: DistilBERt [12]. They found that this stacking approach outperforms each of the individual models (SVM and DistilBERT) when testing on two heterogeneous corpora.

Moreover, according to Chen et al. [3], because each domain has its unique features, including the highly technical financial sector, there is no single magic solution for argument mining to address all domain-specific concerns. Even in the financial domain, there are a variety of data sources, such as 10-K, 10-Q, 8-K, ECC, news articles, and social media.

Argument mining in social media is one of the solid challenges for researchers. One of the reasons is that the limitation of characters count leads to a lack of premises, namely only the claim but no supporters [3]. Hence, we cannot call it an argument, which makes it impossible to apply argument mining procedures. Indeed, the more the document is formal and well-structured is it, the better quality of arguments exist. Therefore, and with respect to financial narratives, earning conference calls seems to be the best candidate.

Table 1: Claim statistics of ECC annotations

	Label	Numerals
Train	In-Claim	1,039
	Out-of-Claim	7,298
Development	In-Claim	114
	Out-of-Claim	1,007
Test	In-Claim	187
	Out-of-Claim	2,196
Total	In-Claim	1,340
	Out-of-Claim	10,571

2.2 Earnings Conference Calls (ECCs)

ECCs are generally held in every fiscal quarter and consist of two main parts; the presentation and the QA session. In the presentation, executives give the statements about the performance of company in the last quarter and exhibit their expectations for the next one. However, the QA session consists of the clarifications, which are asked by the analysts.

Unlike presentation sections, the QA sections are nonstructural since the analysts can ask questions of their will. Although the ECC is an event, transcripts are published on different websites such as Seeking Alpha [1]. As we discussed in the introduction section, the research is based on two approaches; transcript-based and transcript with vocal features. [5, 7, 8, 13] compare models with the same dataset, including the vocal features. The main idea of vocal feature is, to detect the anomaly in the speech according to phonetic rules. Even though the sentence is perfectly structured, if the manager does not believe what he/she says, the model can detect anomaly with those features. Pazienza et al. [14] studied the change in the analysts' reports in the ECCs with an abstract argument mining. That is one of the closest research with FinNum-3. However, they roughly consider each paragraph in only the QA part as one argument, so they did not apply fine-grained argument mining methods.

Besides, the main goal of most of the research is volatility prediction. One of the reasons might be the relatively more straightforward prediction, hence having better results than the exact price detection.

3 MANAGERS CLAIM DETECTION

In FinNum-3, there are two datasets: Investor's claim detection (Chinese) and Manager's claim detection (English). However, since we are not familiar with the Chinese language, we participate only for the English task. Namely, a binary classification problem of in-claim and out-of-claim numerals in the text.

With respect to this English part of data [15], for each data register, we have a complete paragraph of some manager's speech, along with a particular target number, and the goal is to determine if this target number is part of a claim sentence (in-claim), or is not (out-of-claim). Furthermore, every target numeral is also classified into seven categories (e.g., monetary, percentage, temporal, quantity, product number, ranking, other). Four of them are further

[1] https://seekingalpha.com/

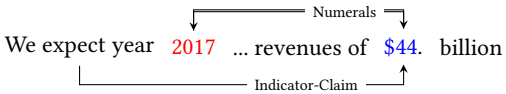

Figure 1: An example of the train dataset [15] including in-claim and out-of-claim numerals.

categorized into several subcategories (e.g., monetary, percentage, temporal, quantity).

The category can be used to train an auxiliary model(same as FinNum-1), to help the final binary classification of the target number. For example, the date category has no in-claim labels that can show us the main structure of the claim-category relations in details. It might be a sufficient input for the joint learning models to predict the claims better. However, we decide not to use the category information in our approach.

Table 1, shows the class distributions for both labels, in the train, development and test splits separately. We can simply observe that this dataset is imbalanced with a ratio of 12.67% (moderate degree of imbalance), which makes the training process more challenging. In addition, different numerals of in-claim and out-of-claims are located even in the same sentence. Therefore, we infer that the sequence of words makes a big difference.

The task organizers chose the evaluation metrics to be the micro-F1 and macro-F1 scores [1], such that the relative contribution of precision and recall are equal.

4 METHODOLOGY

Figure 1 shows an example of the train dataset [15] that include more than one numeral in the same paragraph. We can see that even the sequential numerals can be classified differently in the same sentence. Thus, we had to take the position of words and the sentence context into account.

While deep learning models can produce embeddings that can serve this goal, they need to get trained on big data, our dataset is relatively small. Therefore, we decide to use a transformer model. Transfer learning essence is the ability to transfer previous learned knowledge from one source domain/task to a target one [16]. Among different transformers, BERT-like models have proved to achieve state-of-the-art results in different NLP tasks.

Therefore, we derived the BERT base model (uncased) [17] as a pre-trained model and fine-tuned the weights with our proposed dataset to have more suitable embeddings. In the reminder of this Section, we explain the steps comprehensively.

4.1 Data Preparation

The data is provided along with the splits of training, development, and test (in three different files). Training and development files include the paragraph, "target num", "category", "offset start", "offset end", "claim", despite the test file excluding the category and claim. The paragraphs might have more than one sentence; as we can understand from the Figure 1, a sentence can contain more than one numeral that even the claims can differ.

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

To understand which words are essential to predict a numeral as a claim, we proposed a Decision Tree Classifier model with the help of claim labels. We, first, separated the paragraphs by sentence. Besides, if more than one claim is labeled differently in the same sentence, we segment them into different sentences, even if it is not applicable for the actual task since the test file has no claim labels. In that way, all the sentences have only one kind of claim that eliminates the misinterpretation of the sentences. After removing, the stop words, converting them to lower case, and lemmatization, we transformed all the remaining words as features in the vectors. With only 20 features, the model itself gave results of 78,91% macro-F1 score and 93,68% micro-F1 score. Surprisingly, all the 93,68% success rate comes from the word "expect" -the root of the Decision Tree. That shows us the importance of the indicators in argument mining.

Therefore, we tried to analyze basic features such as; sentence order in the paragraph, sentence length, count of positive and negative words, polarity and subjectivity level. Nonetheless, those extracted features were not helpful at all, as we noticed from the Decision Tree. Basically, the results were the same as before.

We tested different models on this sentence-segmented data, such as SVM, Naive Bayes, CNN, and BERT base models. BERT base model outperformed the other models with 89% macro-F1 and 96,5% micro-F1 scores. The results of the development data can be seen from the Table 4. However, even-though this segmentation is not applicable for the test data, the results were promising, so we understood that sentence-based tokenizing is helpful for the BERT, which is the best performing model. Thus, instead of having one whole paragraph from the Manager's speeches as an input, we split it into single sentence, which includes the target number. So, some of the inconsequential parts are eliminated. Moreover, the primary purpose of the pre-processing is to prepare the data appropriate for the BERT model, and the BERT model deals with the stopwords, lemmatization, etc., by itself. Nevertheless, we cannot give which numeral is in-claim and which one is out-of-claim as a feature to the BERT easily. As the two differently labeled numerals belong to the same sentences, we needed to clarify for which numeral it is labeled. For example, for the sentence in Figure 1, we needed to send the sentence two times; the first one for the "2017" with the label out-of-claim and the second one for the "$44" with the label in-claim. We solve this problem by adding a unique string: **[SPEC]**, as a marker flag, before the target numeral. Also, when we look at the principle of the BERT model, it uses the [MASK] and [CLS] for the prediction. Thus, our added unique marker helps BERT to distinguish the numerals. Figure 2 shows an example of the BERT input representation for classifying the target numeral $44 from the example shown in Figure 1.

Afterwards, we send all the sentence-based tokens, including the markers, to the BERT model for final prediction.

4.2 Model and Training

The concept behind transfer learning models is that the knowledge gained while solving one problem can be useful to a different but related problem. This concept is derived from the human accumulative process of learning. Since a language common knowledge is always appreciated, transformers have rapidly become the model of

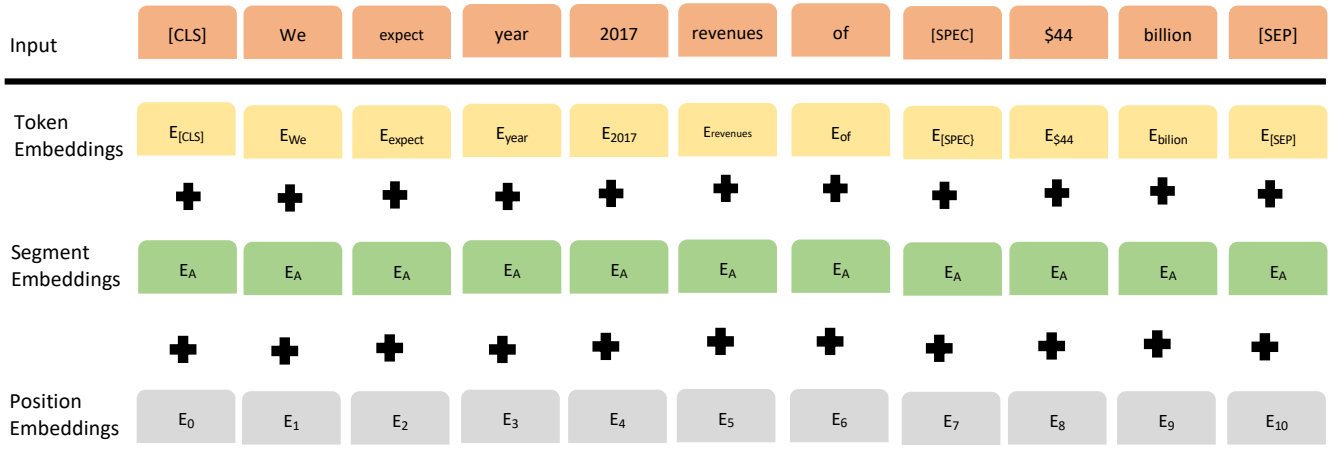

Figure 2: BERT input representation for classifying the target numeral $44 from the example shown in Fig. 1. [SPEC] is a unique string used to distinguish the target numeral from the other tokens.

choice for NLP problems. In particular, the BERT model, which has been trained on a corpus that consists of 16GB data (approximately 3.3 Billion words) from books corpus and Wikipedia. This large corpus of diverse topics enables it to show robust performance for domain-shift problems [18].

One of BERT's other advantages is that the sequence of the words is still kept, which is crucial for us. Indeed, there are various BERT-like models, such as DistilBERT [19], which is 40% less in size, faster to train and retains 97% of the language understanding capabilities of the base BERT model. However, since even the tiny amount of accuracy change is essential for us, with no time or space constraints, we chose to employ the BERT base model. Even though it works significantly slow, such a small-scaled training dataset, does not affect the fine-tuning time seriously.

As we can in Figure 2, BERT model uses three types of embeddings: the Token Embeddings, which are the pre-trained embeddings for different words. On the other hand, the Segment Embedding is to distinguish and sort the additional sentences, and last, the Position Embedding is helpful to determine position of the token in a sentence. Those embeddings are one of the main reasons for the incredible performance and speed of the BERT model. The sentence in Figure 1 demonstrates one of the most challenging predictions for claim classification. The indicator "expect", as we discussed before, refers to in-claim; however, the next numeral after the keyword is "2017", which is labeled as out-of-claim. With the help of Position Embedding, the BERT model understands that the indicator directs to the other target numeral, "$44".

We used the BERT uncased implementation from Huggingface[1]. The used parameters for our fine-tuned BERT model for the sentence classification are stated in Table 2. Several trials to fine-tune the model with changing the learning rate and the batch Size gave us the best configurations.

[1]https://huggingface.co/bert-base-uncased

Table 2: The used parameters to train BERT on our binary classification.

Parameter	Value
Weight Initialisation	Bert-base (Uncased)
Optimizer	Adam
Batch Size	48
Warmup Proportion	0.1
Learning Rate	2e-5
Total Epoch	7
Loss	Cross Entropy

5 EVALUATION

The results are calculated according to F1 scores with average value of macro and micro. The macro parameter on Sklearn represents the metrics calculation for each prediction and finding their unweighted mean. In that way, the comparison of the results is more reasonable.

Table 3: Confusion matrix of the model evaluation on the test data

		Actual	
		In-Claim	Out-of-Claim
Predicted	In-Claim	154	33
	Out-of-Claim	62	2134

The confusion matrix of our results on the test data is represented in Table 3. As we can see our model is able to retrieve most of the True Positive (TP) and True Negative (TN) samples. Yet, out-of-claim prediction reports a small error rate of 1,52%, while in-claim predictions is inadequate with 28,72%. Rather than the test data imbalance, the samples of in-claim in the training data was relatively less. With more data, that cover a fair representative of the test samples, the model could be learned better.

Table 4: Experiment results on the sentence-segmented development data

Models	F1-macro	F1-micro
SVM	79.89	93.83
Naive Bayes	77.56	93.39
CNN	75.86	91.43
Decision Tree	78.91	93.68
BERT base	89.15	96.51

Table 5: The distribution of our false predictions according to category information. False Negative represents the Actual data is in-claim while the prediction is out-of-claim, can be also distinguished by the colors match with Table 3

Category	False Positive	False Negative
quantity_absolute	5	24
quantity_relative	1	3
money	1	13
change	1	2
relative	23	10
absolute	2	9
other	0	1

For a deeper analysis, Table 5 illustrates the category-based error distribution. The false prediction in the category of quantity absolute represents 38% of total False Negative predictions. Likewise, the count of total errors in the relative category depicts 34% of the total errors. Namely, the accurate prediction of those categories might help have better results in the numeral claim detection, although we did not involve in this subtask.

According to [20], the evaluation of all the participants submitted approaches varies in the range of **[85.10%, 97.27%]** for micro-F1 score and in the range **[57.36%, 91.03%]** for macro-F1 score. We are ranked at the near top of the best approaches by a macro-F1 score of **87.12%** and micro-F1 score of **96.01%**. As can be seen from the results, our model can be considered strongly promising.

6 DISCUSSION AND FUTURE WORK

This paper represents our submitted approach to participate at the FinNum-3 task. We proposed a BERT-based model to detect the numerals in the English part of the challenge, namely, Manager's claim in Earnings Conference Calls. The 87.12% macro-F1 score of our fine-tuned BERT model leads us to get the fourth rank, even though the category information is not utilized as an auxiliary model. Despite the fact the our result are promising with respect to other teams, we believe that with the help of extra information and a joint model might improve the overall result. Other learning types may be used to overcome the small size of data like ensemble methods and semi-supervised learning. Moreover, we think that the evaluation strategy should follow more weighted metrics to take into account the imbalance between the two labels.

REFERENCES

[1] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 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*, pages 19–27, 2019.

[2] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-15 finnum-2 task: Numeral attachment in financial tweets. *Development*, 850(194):1–044, 2020.

[3] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. *From Opinion Mining to Financial Argument Mining*. Springer Nature, 2021.

[4] Katherine A Keith and Amanda Stent. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. *arXiv preprint arXiv:1906.02868*, 2019.

[5] Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. Voltage: volatility forecasting via text-audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013, 2020.

[6] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451, 2020.

[7] Yu Qin and Yi Yang. What you say and how you say it matters: Predicting financial risk using verbal and vocal cues. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, page 390, 2019.

[8] Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070, 2020.

[9] John Lawrence and Chris Reed. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818, 01 2020.

[10] Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. A stacking approach for cross-domain argument identification. In *International Conference on Database and Expert Systems Applications*, pages 361–373. Springer, 2021.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[13] Zhen Ye, Yu Qin, and Wei Xu. Financial risk prediction with multi-round q&a attention network. In *IJCAI*, pages 4576–4582, 2020.

[14] Andrea Pazienza, Davide Grossi, Floriana Grasso, Rudi Palmieri, Michele Zito, and Stefano Ferilli. An abstract argumentation approach for the prediction of analysts' recommendations following earnings conference calls. *Intelligenza Artificiale*, 13(2):173–188, 2019.

[15] 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*, pages 1973–1976, 2020.

[16] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[18] Minho Ryu and Kichun Lee. Knowledge distillation for bert unsupervised domain adaptation. *arXiv preprint arXiv:2010.11478*, 2020.

[19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[20] 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.