NumClaim: Investor's Fine-grained Claim Detection

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
QF, NTHU, 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

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

The goal of claim detection in argument mining is to sort out the key points from a long narrative. In this paper, we design a novel task for argument mining in the financial domain, and provide an expert-annotated dataset, NumClaim, for the proposed task. Based on the statistics, we discuss the differences between the claims in other datasets and the claims of the investors in NumClaim. With the ablation analysis, we show that encoding numeral and co-training with the auxiliary task of the numeral understanding, i.e., the category classification task, can improve the performance of the proposed task under different neural network architectures. The annotations in the NumClaim is published for academic usage under the CC BY-NC-SA 4.0 license.

CCS CONCEPTS

• **Information systems → Information extraction**.

KEYWORDS

Argument mining, claim detection, joint learning

ACM Reference Format:
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 and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412100

1 INTRODUCTION

Argument mining has attracted much attention recently, and some related datasets are released for different topics [3]. However, none of the previous work analyze the argument mining issue in the financial domain. In this paper, we explore the task that detects the claims from the reports written by professional stock analysts.

There are many publicly available professional and amateur investors' analysis reports on the websites of the investment banks and the financial social media platforms, respectively. In an analysis report, the investors first write down their stances toward a certain financial instrument, i.e., bullish or bearish, and then provide several claims to support their stances. Most of the previous work focus on the stances in the analysis reports only [14]. Some recent work provide an in-depth analysis of the fine-grained information such as the price target of the professional and amateur investors [6, 15]. However, none of the above related work detect the claim in the analysis report. In this paper, we provide the first dataset, called NumClaim, for the claim detection in financial analysis reports.

Table 1: Instances in NumClaim. The bold numeral is the target numeral.

Sentence	Label
We estimate that the sales growth rate may exceed **40**%.	In-claim
Professional audio/visual products account for **20**%.	Out-of-claim

Numerals provide important information in financial narratives [6, 7]. Our statistical result in the financial analysis reports shows that over 58.47% of sentences contain at least one numeral. Without the numerals, lots of fine-grained information in the analysis reports will be lost. This phenomenon evidences the importance of the numerals in the financial narrative. We will detail the statistics in Section 3.

Taking a close look at the analysis reports, we further find that investors always make a claim with an estimation. The first sentence shown in Table 1 is an example. When investors make a claim in the analysis report, they will provide a fine-grained estimation such as "the sales growth rate may exceed **40**%". This kind of estimations provide more fine-grained opinions than only a stance (bullish or bearish). For example, comparing with the first sentence in Table 1, the claim, "the sales growth rate may exceed **80**%", is stronger. As shown in Table 1, the second sentence containing the numeral "20" cannot be interpreted as the investor's claim. Instead, it just describes the proportion of audio/visual products.

In this paper, we focus on the fine-grained claims of the professional stock analysts, i.e., the claims accompanied with numerals in the financial analysis reports. That is, given a sentence in the financial analysis report and a target numeral in this sentence, we attempt to predict whether the given numeral is an in-claim numeral or not.

Our contributions are threefold: (1) We explore the argument mining issue in finance, and design a novel task suitable for the nature of investors' narrative. (2) We provide an expert-annotated

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412100

dataset, NumClaim[1], for the proposed task. (3) We show that encoding with numeral encoder and co-training with the numeral understanding auxiliary task are helpful for the numeral-oriented task.

We will discuss the following research questions in the rest of this paper.

- **(RQ1)** What are the differences between the claims in other sources and the claims of investors?
- **(RQ2)** To what extent can we improve the performance of the proposed task by incorporating numeral information?
- **(RQ3)** Whether joint learning with numeral understanding task works for the proposed task?

2 RELATED WORK

Recently, argument mining is one of the hot topics in the natural language processing (NLP) community [7, 13, 19, 21, 24]. Claim detection is the first step when researchers attempt to analyze the argument data. Previous work mainly focus on the Wikipedia, customer review, and debate data [1, 2]. Few previous work discuss the arguments in the financial domain. In this paper, we first provide an in-depth analysis of the professional analysis reports, and attempt to detect the investor's claim from these reports.

Analyzing the numeral information in documents is an emerging topic, and attracts much attention [7, 18, 23, 26, 29]. Based on our observation, investors always write down a claim with an estimation, which is represented as a numeral. Some previous works [4, 8] show that these estimations are informative for investment decisions. Leveraging to this special narrative style, we design a numeral encoder and a numeral understanding auxiliary task to enhance the ability of neural network models toward investor's claims. We adopt the taxonomy in our previous work [6] to annotate the category of the target numeral, and further experiment with the numeral understanding auxiliary task. The experimental results evidence that the proposed methods are useful for the numeral-oriented task.

3 DATASET

3.1 Dataset Construction

We collect the Chinese financial analysis reports written by the professional stock analysts from Bloomberg, and ask annotators to label the category of a given numeral and further tell if it is an in-claim numeral. The annotators work in the financial industry (bank's treasury department and hedge fund). Because the numeral taxonomy in Chen et al. [6] is designed for financial social media data, some subcategories related to technical analysis index are not mentioned in the analysts' reports. We modify the numeral taxonomy for analysts' reports. The categories and the subcategories are listed in Table 2. Two experts working in the financial industry annotate the category label and the claim label of the given numeral. If the labels annotated by these two experts are different, the third annotator will be invited to confirm one of the annotated labels. The Cohen's kappa agreements [10] between the experts are 89.55% and 88.31% for the category label and the claim label, respectively. Note that, in this paper, we follow previous works [6, 26] to focus on the numerals represented by digits.

[1]http://nlg.csie.ntu.edu.tw/nlpresource/NumClaim/

Table 2: Statistics of NumClaim.

Category	Subcategory	In-claim	Out-of-claim	Sum
Monetary	price	42	33	75
	money	506	368	874
	change	3	15	18
Percentage	absolute	208	500	708
	relative	408	402	810
Temporal	date	0	2,134	2,134
	time	0	3	3
Quantity	absolute	55	219	274
	relative	0	4	4
Product Number		1	135	136
Ranking		0	3	3
Other		0	105	105
Total		1,223	3,921	5,144

Table 3: Statistics of argument mining datasets.

Dataset	NumClaim	CRC [13]	PE [12]
Language	Chinese	Chinese	English
Source	Analysis Report	Hotel Review	Persuasive Essay
# Word	42,594	21,848	97,420
# Numeral	5,144	67	111

Table 2 shows the statistics of NumClaim. Among 5,144 instances in the NumClaim dataset, 23.78% and 76.22% of instances containing numerals are annotated as "In-claim" and "Out-of-claim", respectively. Some categories such as "Temporal" may not be labeled as a claim in the analyst's report. We separate 80% of the data as the training set, the rest 20% of the data are used as the test set, and take 10% of the training data as the development set.

3.2 Comparison between Datasets

In this section, we will provide an answer to **(RQ1)** by comparing the proposed dataset, NumClaim, with a Chinese argument mining dataset [13] and an English argument mining dataset [12] covering hotel review and persuasive essays, respectively. Table 3 shows that numerals play an important role in financial analysis reports, and occupy a higher proportion than those datasets from other sources.

The other difference between the investors' claims and the claims in hotel reviews or persuasive essays is that investors make claims on future events, such as the price movement or the estimated earning. In contrast, the claims in hotel reviews and persuasive essays describe past experiences and facts. That is, investors' claims are a kind of prediction based on the latest market information, and the claims in hotel reviews or persuasive essays are the opinions based on the writer's experiences.

We further adopt Chinese Readability Index Explorer [27] to compare the readability of Chinese argument mining datasets. Table 4 shows the results. Since the analysis reports are more domain-specific than the hotel reviews, more difficult words (the words out of the top 8,000 frequent Chinese terms, which cover 95% of general corpora) appear in NumClaim. The statistics also show that analysts use a few negative words when writing the reports. Analysts

Table 4: Readability analysis of Chinese datasets.

	NumClaim	CRC
# hard words	**31.95**	18.28
# negative words	0.14	**0.60**
# synonym	0.28	**1.49**
Noun phrase modifier ratio	0.29	**0.38**
Noun phrase ratio	**31.79**	26.62
# transition words	**4.86**	1.62

Table 5: PMI scores of the words in analysts' reports.

In-claim		Out-of-claim	
estimate	2.86	lower/higher than	-1.37
price target	2.80	cause	-1.37
downgrade	2.58	last year	-1.26
upgrade	1.55	influence	-1.25

avoid using synonyms because lots of synonyms may cause reference problems and make the articles hard to follow. Using noun phrase modifiers may lower the readability. Comparing with the hotel reviews written by the crowd, analysts use less noun phrase modifiers. The ratio of noun phrases can be used to evaluate the informativeness in the sentence. We find that sentences in analysts' reports are more informative than those in hotel reviews. Besides, analysts use more transition words transition words such as "firstly" and "therefore", which are important keys to argument detection, than the crowd.

3.3 Comparison between the Context of In-claim and Out-of-claim Numerals

In this section, we provide comparisons between the context of the numeral in a claim and the context of the numeral not in a claim. Based on the analysis results of Chinese Readability Index Explorer [27], there are no difference between the context near to different labels from readability aspect. We further calculate the pointwise mutual information (PMI) of the words near to different labels. PMI is one of the popular measures for constructing a lexicon for classification tasks such as sentiment analysis [22, 28]. We follow the previous work [20] to calculate the PMI score by subtracting the PMI(w, "Out-of-claim") to PMI(w, "In-claim"), where w denotes the target word. Table 5 shows the statistics. The words representing subjective opinions are always near to in-claim numerals, and the words describing the cause and effect or the comparison are always near to out-of-claim numerals.

4 MODELS

Given a sentence in the analysis report and a target numeral in this sentence, the model will predict whether the given numeral is an in-claim numeral or not. We explore the models by using different features and different techniques. We employ the pre-trained BERT [11] to encode the sentence and use the last layer as the representation. That is, we use a $768 \times w$ tensor to represent a sentence with w characters. We adopt Adam optimizer [17], and use both early-stop with five-epoch patience by monitoring the loss of development set and dropout layer (0.3) to prevent overfitting. The details of the models are described in the following subsections.

4.1 Vanilla Neural Network Architecture

We adopt vanilla convolutional neural network (CNN), bidirectional gated recurrent unit (BiGRU), and capsule network (CapsNet) architectures as our baseline models. In CNN model, we use one convolutional layer with a max-pooling layer [16], and a multi-layer perceptron to make the binary classification between "in-claim" and "out-of-claim" numerals. BiGRU model consists of a BiGRU layer with a multi-layer perceptron [9]. In the CapsNet architecture [5], we extract the features by the CNN layer and shrink the features by squashing function [25]. Since the labels are imbalanced, we use the reciprocal of the proportion to set the class weight (CW), i.e., 4.2052 and 1.3120 for both "in-claim" and "out-of-claim" numerals, respectively.

4.2 Numeral Encoder (NE)

We represent the target numeral digit-by-digit and add a magnitude embedding to present the intra-numeral position information. That is, for each digit in the target numeral, we use an 11×1 tensor to represent the digit (0–9) and the decimal point, and concatenate the digit embedding with a $n \times 1$ tensor for the inter-numeral position information. In this paper, we set n to 7. In the experiments in Table 6, we use the CNN architecture to encode the numeral information, and then concatenate the numeral information with the context information.

4.3 Joint Learning with Category Classification Task

Multi-task learning can enrich the information during the training process [5, 11]. One of the important numeral understanding tasks is category classification. For example, the numeral "1996" in "This company established in **1996**" is labeled as "date". As shown in Table 3, none numerals in "date" category are labeled as an in-claim numeral. Therefore, in order to enrich the numeral understanding ability of our models, we add an auxiliary task that classifies the category of a numeral. In addition to providing the claim labels, the models are asked to predict the category of the target numeral, i.e., classifyingthe target numeral into 12 subcategories listed in Table 2.

5 RESULTS AND DISCUSSION

In this section, the word "significant(ly)" means the results are significantly different at $p < 0.05$ using McNemar's test. Table 6 shows the results of different models. Full models, i.e., the models with class weight (CW), numeral encoder (NE), and category classification task (CG), are significantly better than that of the baseline models, i.e., those without class weight, numeral encoder, and joint learning setting. These results indicate that encoding the symbolic information of the target numeral, i.e., numeral encoder, and the semantic understanding task of the target numeral, i.e., category classification task, are useful in the proposed task when training a neural network model. In other words, both symbolic information and the numeral category are not learned in the BERT representation originally.

Table 6: Experimental results.

Architecture	CNN	BiGRU	CapsNet
Baseline	76.15%	77.97%	77.93%
+ CW	77.26%	78.29%	78.68%
+ CW & NE (CNN)	78.19%	79.06%	80.91%
+ CW & NE (CNN) & CG	81.35%	81.65%	**82.62%**

With the ablation analysis, we also find that using class weight improve the performance of the imbalance data. The results between the models with and without the CW setting are significantly different under the BERT-CapsNet architecture, but the results are not statistically significant between other architectures. Based on the ablation analysis, the models with numeral encoders perform significantly better than the baseline models. These comparisons give a positive answer to **(RQ2)**, and show that adding numeral encoder can increase the macro-averaged F1-score by 3.48% in the proposed task.

We further compare the results of the models co-training with the auxiliary task. We find that the numeral encoder and the numeral understanding task provide different information to the models. Thus, when adopting both schemes, the performances are enhanced under different neural network architectures. These results also show a positive answer to **(RQ3)**. That is, joint learning with numeral understanding task works for the proposed task.

Furthermore, we find that the model with CNN performs better in providing information about the target numeral. The Macro-F1 scores of BERT-CapsNet (+ CW & NE & CG) with CNN, BiGRU, and CapsNet numeral encoders are 82.62%, 79.77%, and 78.11%, respectively.

Besides, we find that models make wrong predictions on some instances containing high PMI scores words. For example, the "8.0" in the description "lower than the original estimation +8.0%QoQ" is not an in-claim numeral, but the reason why the models make a wrong prediction on this case may due to the word "estimation".

6 CONCLUSION

In this paper, we explore the argument mining issue in the financial domain, and propose a high-quality expert-annotated dataset, NumClaim, under the CC BY-NC-SA 4.0 license. Based on experimental results, we show the difference between the investors' claims and the claims in other datasets. With the strict comparisons, we find that learning with the numeral encoder and the auxiliary task of category classification can improve the performance on investor's claim detection task.

In the future, we plan to extend the argument mining in finance to both premise detection and relation linking between the claim and the premise. The comparison between the claims from the investors with different stances is also an important issue. Furthermore, the rationality assessment of the claim and its premise is one of the important challenges when evaluating the fine-grained opinions of the investors.

ACKNOWLEDGMENTS

This research 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] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A Benchmark Dataset for Automatic Detection of Claims and Evidence in the Context of Controversial Topics. In *Proceedings of the First Workshop on Argumentation Mining*.
[2] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance Classification of Context-Dependent Claims. In *EACL*.
[3] Elena Cabrio and Serena Villata. [n.d.]. Five years of argument mining: a data-driven analysis. In *IJCAI*.
[4] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. [n.d.]. Crowd View: Converting Investors' Opinions into Indicators.. In *IJCAI*.
[5] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. [n.d.]. Numeral attachment with auxiliary tasks. In *SIGIR*.
[6] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. [n.d.]. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *WI*.
[7] 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 *ACL*.
[8] Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. [n.d.]. Crowdpt: Summarizing crowd opinions as professional analyst. In *WWW*.
[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv:1406.1078* (2014).
[10] Jacob Cohen. [n.d.]. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 ([n. d.]).
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.
[12] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *ACL*.
[13] Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!. In *COLING*.
[14] Colm Kearney and Sha Liu. [n.d.]. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33 ([n. d.]).
[15] Katherine Keith and Amanda Stent. [n.d.]. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. In *ACL*.
[16] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*.
[17] Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
[18] 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 *EMNLP*.
[19] Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *COLING*.
[20] Quanzhi Li and Sameena Shah. 2017. Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits. In *CoNLL*.
[21] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *ACL*.
[22] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. [n.d.]. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *SemEval*.
[23] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring Numeracy in Word Embeddings. In *ACL*.
[24] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. In *EMNLP*.
[25] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. [n.d.]. Dynamic routing between capsules. In *NeurIPS*.
[26] Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *ACL*.
[27] Hou-Chiang Tseng, Berlin Chen, Tao-Hsing Chang, and Yao-Ting Sung. 2019. Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering* 25 (2019).
[28] Duy Tin Vo and Yue Zhang. 2016. Don't Count, Predict! An Automatic Approach to Learning Sentiment Lexicons for Short Text. In *ACL*.
[29] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *EMNLP*.