# TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data

Jose G. Moreno
jose.moreno@irit.fr
University of Toulouse - IRIT
F-31000, Toulouse, France

Emanuela Boros
emanuela.boros@univ-lr.fr
University of La Rochelle - L3i
F-17000, La Rochelle, France

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle - L3i
F-17000, La Rochelle, France

## ABSTRACT

This paper presents the TLR participation in the FinNum-2 task. Our system is based on a Transformer architecture improved by a pre-processing strategy for numeral attachment identification. Instead of relying on a vanilla attention mechanism, we focus the attention to specific tokens that are essential for the task. The results in an unseen test collection show that our model correctly generalises the predictions as our best run outperforms all those of other participants in terms of F1-macro (official metric). Further, results show the robustness of our method as well as the experiments with two alternatives (with and without parameter tuning) leading to an additional improvement of 4% over our best run.

**ACM Reference Format:**
Jose G. Moreno, Emanuela Boros, and Antoine Doucet. 2018. TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

## TEAM NAME

TLR

## SUBTASKS

Numeral attachment in financial tweets (English)

## 1 INTRODUCTION

Social media platforms are becoming a main source of information nowadays [5]. News media, politicians, personalities, etc., use microblogs such as Twitter daily to briefly communicate with their target public (followers). As an example, the current President of the United States of America[1] publishes an average of seven to ten tweets per day, approximately totalling 3,000 tweets per year to an audience of 86 millions of followers [9, 10].

However, politicians are not alone in the use of social media. Companies also use social media to publish information about their current and new products, successful histories, or information to their shareholders, including financial information. Similarly, shareholders or the general public also share financial information on social media through the use of special identifiers [2]. Indeed, social media platforms use some special characters to allow users to tag information. One of the commonest is the hashtag token (#) that is used to tag conversations in Twitter [11]. Another special character is the at symbol (@) that is usually used to mention users of the platform. The main difference is that the latter refers to unique elements while the former may be related to ambiguous topics. A similar token of the latter group is the dollar symbol ($) used on the financial-related information. It is widely promoted by Cashtag[2] platform. This platform is based on the use of $cashtags that are unique identifiers for individuals and businesses using Cash App. The $Cashtags allow an aggregation of the information related to a unique organisation as they are used as identifiers.

The use of $Cashtags opens promote the exploration of current challenges and techniques in information extraction (IE) applied to the financial domain. In this context, multiple natural language processing (NLP) tasks [2–4, 7] can be addressed automatically to improve user experience when using $cashtags or to mine vital information from their use. This includes named entity recognition and linking, relation extraction and classification, or numeral attachment identification and classification, to mention a few of tasks that may be associated with the use of $cashtags. During its 15th edition, NTCIR hosted the numeral attachment challenge, where participants are asked to automatically identify the relatedness of numeral information and $cashtags within financial tweets. However, several recent models in IE tend to give hardly any attention to this type of information. FinNum-1 [3] and FinNum-2 [4] addressed the problem of the understanding of numbers in financial information, where fine-grained numeral understanding in financial social media data is essential to link $Cashtags and numeral data.

In this paper, we present the TLR participation in the FinNum-2 task. Our system is based on recent architectures based on neural language models and a simple but effective explicit attention mechanism. Our best official run outperforms other participants on the task with a significant margin. Moreover, improvements over our own runs can be obtained by the use of an ensemble strategy but on a larger number of predictions.

The remainder of this paper is organised as follows. Section 2 presents the background information related to the task, the works that inspired our model and the details of our models. The experimental setup, our official results, and complementary experiments are elaborated in Section 4. Finally, the conclusions are drawn in Section 5.

[1] These statistics are based on the @realDonaldTrump Twitter account.

[2] https://cash.app/


*Woodstock '18, June 03–05, 2018, Woodstock, NY*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

## 2 BACKGROUND

In this section, we briefly introduce recent neural models based on Transformers [14], such as BERT and RoBERTa, and their use for relation extraction (RE). Our intuition is that the RE task is closely related to the numeral attachment task.

### 2.1 Neural-based Language Models

Given the strong performance of recent deep architectures trained on variants of language modelling, we chose BERT [6] and RoBERTa [12] models. These architectures have been successfully evaluated in a wide number of NLP tasks [1, 6, 13, 15, 16].

Both use the same architecture based on several layers of the Transformer blocks. A Transformer block [14] is a deep learning architecture based on multi-head attention mechanisms with sinusoidal position embeddings[3]. It is composed of a stack of identical layers, each layer having two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by a layer normalisation. All sub-layers in the model, as well as the embedding layers, produce outputs of dimension 512.

The RoBERTa model is based on different modifications of the BERT pre-training procedure that improve end-task performance.

A binary text classification based on the Transformer (either BERT or RoBERTa) is depicted in Figure 1.

### 2.2 Relation Extraction with Transformers

Extracting relations between tokens in sentences is a challenging NLP task. However, a recent work on relation classification Baldini Soares et al. [1] showed that vanilla Transformer-based[4] sequence classifiers are strong enough to identify relations. This is achieved by the introduction of additional markers that help the model to drive their attention mechanics. Indeed, Transformer-based models are already strong to classify sentences. However, it may struggle when the same entry is considered for multiple classes (positive and negative). To address this problem, Baldini Soares et al. [1] proposed a pre-processing step that is required to indicate entity tokens by using extra tokens in the input sentence. Then, the typical sequence classification strategy proposed by Devlin et al. [6] can be used. This strategy consists of using a feed-forward layer that takes the [CLS] token representation and that is trained to perform the classification task. This simple but powerful architecture is privileged in our work.

## 3 NUMERAL ATTACHMENT IN FINANCIAL TWEETS

Although detailed information regarding the task can be found in Chen et al. [4], we briefly describe hereafter the task.

[3]In practice, these models use absolute positional embeddings [8] instead, as a common practice.

[4]The BERT model is used in [1].

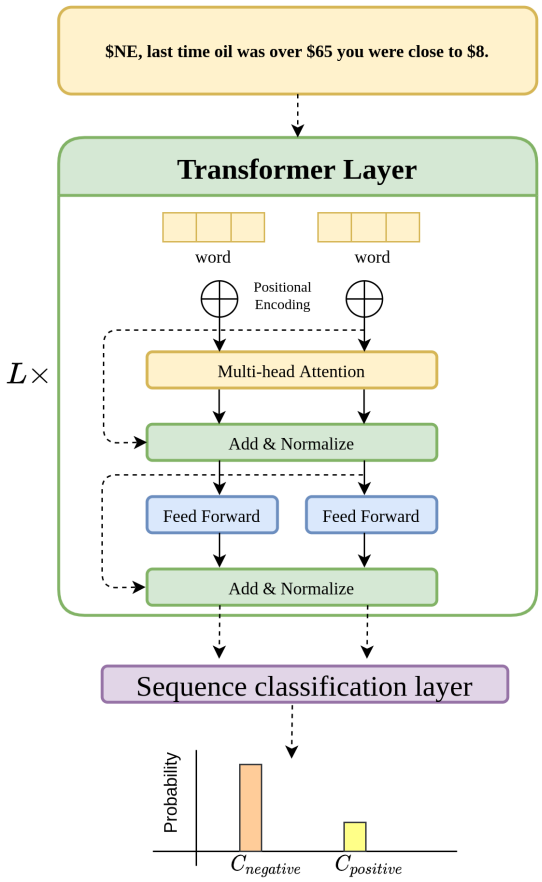


**Figure 1: BERT architecture for binary text classification. L is the number of Transformer layers. In our experiments, the BERT and RoBERTa models have 12 Transformer blocks.**

### 3.1 Task Definition

Given a tweet (in text format), a *$cashtag* offset, and a numeral offset, the FinNum-2 task consists in determining whether a numeral indicated by the numeral offset relates or not to the *$cashtag* indicated by the *$cashtag* offset. Two examples of this task are presented in Figure 2 (a), where the $NE token is positively attached to $8 and negatively to $65. Note that in this case the same sentence is associated with two examples depending on the attached numeral. Thus, this is a binary classification task. In the context of the NTCIR, three files are shared by the organisers depending on the task stage. In particular, the train and development sets are provided with labels at the beginning of the task while the test set is provided without labels. Finally, after the official results are published, the labels of the test set are shared with all participants.

### 3.2 Numeral Attachment Classification

We opted for a Transformer neural-based language model (as described in Section 2) and we introduced a pre-processing technique for the numeral attachment task. In our case, we mainly focus on

the input preparation to facilitate the system identification of the key information for the task.

Regarding the necessary pre-processing step, we first add two reserved words to mark the beginning and the end of the *$cashtag* mentioned in the text. We introduce the **£** and **§** additional reserved tokens, and we mark the words concerned in the sequence. Figure 2 presents the (a) initial input provided by the organisers, (b) the transformed information for our system. Note that the transformed information transcribed in the text formats only the tokens concerned in the classification (Figure 2 (c)).

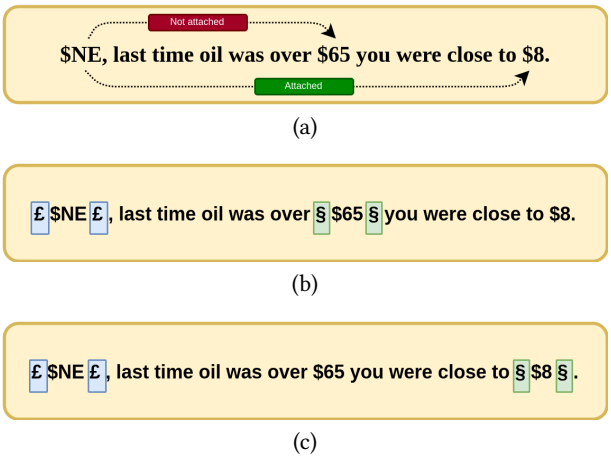


**Figure 2: Input pre-processing step with marked key information.**

Our predictions for each model are based on the output probability obtained by the model as depicted in Figure 1. Finally, we perform an ensemble strategy based on a *max* or *min* selection to define the final class prediction.

## 4 EXPERIMENTS AND RESULTS

### 4.1 Dataset

A manually annotated dataset was provided by the organisers of the FinNum-2 task. This dataset is composed of 7,187 training examples, 1,044 validation examples, and 2,109 test examples. The details of the dataset and the manual annotation process can be found in Chen et al. [2, 4].

### 4.2 Metrics

The official metric of the task is the F1-macro computed as the harmonic mean between precision and recall.

### 4.3 Analysis on the Validation Partition

Despite the multiple parameters involved in the architecture based on Transformer layers, we opted for a standard configuration of the models as shown in Table 1.

The main parameter that was selected using the validation partition is related to the number of epochs used to train the model. We explored a total of 20 epochs and selected the model with the highest validation performance between the epochs. Results for BERT and RoBERTa models are presented in Figure 3. Best performances are obtained in epoch 2 and 4 for BERT and RoBERTa, respectively. From Figure 3, we can also see that (1) for both models, the choice of three epochs seems an inadequate option so this shows the relevance of this parameter, (2) later epochs (after 15) the performances between BERT and RoBERTa are indiscernible.

**Table 1: Parameters used for our BERT and RoBERTa models.**

| Name | Value |
|---|---|
| Weight Initialisation | BERT-base / RoBERTa-base |
| Batch Size | 32 |
| Optimiser | Adam |
| Learning Rate | $3 \times 10^{-5}$ |
| Epsilon | $10^{-8}$ |
| Clipnorm | 1 |
| Loss | Sparse Binary Crossentropy |

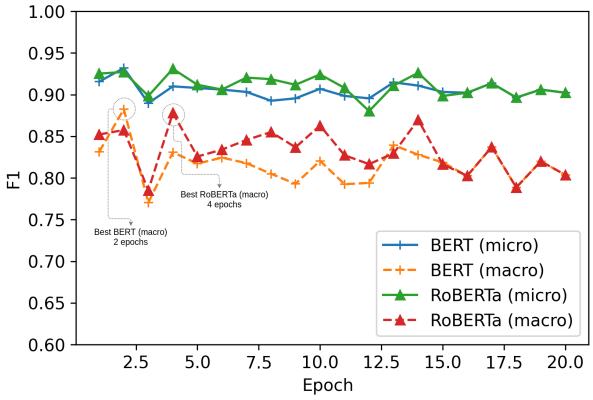


**Figure 3: Performance of the model over the validation partition.**

Additionally, our ensemble strategy consists in selecting the *max* or *min* function as the predictor for our last run. Results are presented on Table 2 for these two functions. In the validation set, the BERT model outperformed the RoBERTa model while both models are outperformed by the *min* function. Thus, these three models were selected for our participation on the task[5].

### 4.4 Official Results

Our team submitted three runs that were calculated as follow:

- Run 1: It consists in the BERT predictions of our model.
- Run 2: It consists in the RoBERTa predictions of our model.
- Run 3: It consists in the *min* between the BERT and RoBERTa predictions of our model.

[5]We discarded *max* function as each participation was limited to three runs and it showed the lowest performance.

**Table 2: Results using the validation partition for *min* and *max* functions.**

| | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.9013 | 0.8667 | 0.8826 |
| RoBERTa | 0.9088 | 0.8542 | 0.8781 |
| *min* | 0.8850 | 0.8926 | 0.8887 |
| *max* | 0.9385 | 0.8284 | 0.8704 |

The official results ordered by run performance (rank) are presented in Figure 4. Our *run 3* (*min*) outperformed all other participants of the task while *run 1* (BERT) and *run 2* (RoBERTa) achieved the 3rd and 6th position in terms of F1-macro, respectively. We can note that our *run 2* is part of a group of runs denoted *group 2*. The runs in the *group 2* have very similar performance values suggesting that there are few chances of observing statistical differences between them[6]. A similar situation can be observed in the case of *group 3*. Indeed, after *group 3* it seems to exist more variation between runs. Based on the observation of low intra variance within *groups 2* and *3*, we intuit that combination of any couple of runs from the two groups may deal with similar improvements. This extra exploration is studied in section 4.5.2.

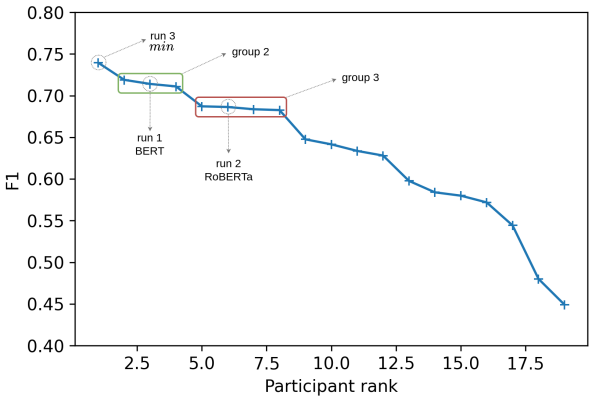


**Figure 4: F1-macro performances ordered by participant rank. Our three runs are identified by dotted circles.**

## 4.5 Unofficial Results

In order to understand the improvement of our model, we perform additional experiments that were not submitted as official runs.

*4.5.1 Impact of min ensemble.* The combination of our BERT and RoBERTa models was successful when using the *min* function. It clearly outperformed the *group 2* results (second best). Moreover, the *min* function can be interpreted as a higher threshold strategy as a positive classified example must be considered positive by our BERT model as well as our RoBERTa model. This can be individually analysed by making it harder for each system to predict an example as positive. So, instead of considering all examples with a probability greater than 0.5 for the positive class, we variate this threshold[7] and presented the results in Table 3. Note that most of the F1 performances increased as the threshold value is higher. However, our BERT and the *min*-based strategy models achieve their best performances at 0.7 and decrease after that. This result suggests that this parameter must be carefully tuned.

**Table 3: Results of our runs using the test partition. Threshold for the positive class was increased from 0.5 (official runs) to 0.9.**

| | | Precision | Recall | F1 |
|---|---|---|---|---|
| 0.5 | BERT | 0.8015 | 0.6793 | 0.7141 |
| | RoBERTa | **0.8435** | 0.6484 | 0.6864 |
| | *min* | 0.8016 | 0.7078 | 0.7395 |
| 0.6 | BERT | 0.7938 | 0.7184 | 0.7461 |
| | RoBERTa | 0.8301 | 0.6612 | 0.6996 |
| | *min* | 0.7866 | 0.7387 | 0.7585 |
| 0.7 | BERT | 0.7731 | 0.7478 | 0.7592 |
| | RoBERTa | 0.8114 | 0.6716 | 0.7081 |
| | *min* | 0.7691 | 0.7632 | **0.7661** |
| 0.8 | BERT | 0.7445 | 0.7648 | 0.7538 |
| | RoBERTa | 0.7796 | 0.6844 | 0.7147 |
| | *min* | 0.7380 | 0.7733 | 0.7528 |
| 0.9 | BERT | 0.7228 | 0.7841 | 0.7438 |
| | RoBERTa | 0.7646 | 0.7115 | 0.7324 |
| | *min* | 0.7169 | **0.7896** | 0.7388 |

*4.5.2 Baselines and extra ensemble combinations.* Following the same configuration setup and parameters, we train multiple extra models to better understand the real improvements of our models (when compared against original models) and a further understanding of the ensemble *min* function:

- A vanilla BERT with any input modification.
- A vanilla RoBERTa with any input modification.
- A *min* ensemble based on our BERT model (three models).
- A *min* ensemble based on our RoBERTa model (three models).
- A *min* between the ensemble *min* of the BERT and RoBERTa predictions.

**Table 4: Unofficial results of our models and baselines using the test partition. We ran three times our models instead of only ones and applied the *min* ensemble function. Parameters remain unchanged w.r.t. our official runs.**

| | | Precision | Recall | F1 |
|---|---|---|---|---|
| Baselines (vanilla models) | BERT | 0.8134 | 0.6638 | 0.7004 |
| | RoBERTa | **0.9182** | 0.6041 | 0.6313 |
| *min* Ensemble (n=3) (our models) | BERT | 0.7933 | 0.6949 | 0.7267 |
| | RoBERTa | 0.8204 | 0.7090 | 0.7447 |
| | *min* | 0.7964 | **0.7489** | **0.7688** |

[6]This is arguable because it depends on the scale. However, it seems fair to assume.

[7]between 0.5 and 0.9

Results for the validation and test partitions are presented in Table 4. Note that aggregating ensemble methods have a beneficial effect in our model. Indeed, the *min* plus *min* function outperforms all the models including our submissions. Note that this result is obtained without extra parameters nor any special tuning. Despite these positive results, we strongly believe that there is still room for improvement as our model is based on a standard sequence classification strategy and more elaborated representation may be included in the model by using not only the [CLS] representation but also the representation of the **£** and **§** tokens.

## 5 CONCLUSIONS

This paper presents our participation in the FinNum-2 task at NTCIR-15. The proposed model is based on an information extraction strategy that combines Transformer-based models with positional information. Our main finding is that this representation is relevant for the task of numeral attachment identification. Our best run achieved the top performance in the F1-macro, the official metric. As future work, we intend to evaluate the quality of the proposed model into other financial tasks such as fine-grained numeral understanding [3].

## ACKNOWLEDGEMENTS

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grant 825153 (EMBEDDIA).

## REFERENCES

[1] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *ACL*. 2895–2905.

[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, 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 NII Testbeds and Community for Information Access Research (NTCIR-14) Conference (NTCIR-14)*.

[4] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 task: Numeral Attachment in Financial Tweets. In *Proceedings of the 15th NII Testbeds and Community for Information Access Research (NTCIR-15) Conference (NTCIR-15)*.

[5] Mary J Culnan, Patrick J McHugh, and Jesus I Zubillaga. 2010. How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive* 9, 4 (2010).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Ismaıl El Maarouf, Youness Mansar, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Fortia Financial Solutions. 2020. The finsim 2020 shared task: Learning semantic representations for the financial domain. In *The Second Workshop on Financial Technology and Natural Language Processing in conjunction with IJCAI-PRICAI 2020*. 81.

[8] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 1243–1252.

[9] Panayota Gounari. 2018. Authoritarianism, discourse and social media: Trump as the 'American agitator'. *Critical theory and authoritarian populism* (2018), 207–227.

[10] Wendy Hall, Ramine Tinati, and Will Jennings. 2018. From Brexit to Trump: Social media's role in democracy. *Computer* 51, 1 (2018), 18–27.

[11] Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. 173–178.

[12] 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).

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[14] 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.

[15] Yu Wang, Yining Sun, Zuchang Ma, Lisheng Gao, Yang Xu, and Ting Sun. 2020. Application of Pre-training Models in Named Entity Recognition. *arXiv preprint arXiv:2002.08902* (2020).

[16] Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2361–2364.