Using BERT to Profiling Cryptocurrency Influencers

(Notebook for PAN at CLEF 2023)

Daniel Yacob Espinosa, Grigori Sidorov

Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico

Abstract

In recent years, the rise of influential individuals in cryptocurrencies on social media has played a significant role both on the internet and in investments. One of the major challenges in this area is determining which users are experts or influencers who can directly contribute to the value of cryptocurrencies. In this instance, PAN 2023 created the task "Profiling Cryptocurrency Influencers with Few-shot Learning" in which we are participating [1]. Our solution approach for this task involves using the BERT model with a preprocessing layer, where we achieved classification results with over 92% precision for all three subtasks included in this challenge.

Keywords

BERT, Cryptocurrency Influencers, Tweets, Twitter

1. Introduction

Since the emergence and popularity of Bitcoin in 2009[2], new ways of conducting commerce have arisen. Similarly, the internet has played a significant role in making this cryptocurrency popular worldwide. In this case, we are primarily referring to social media networks. As we have seen in recent years, social media has been a very useful mechanism for communication, whether it's for viralizing content such as videos or photos, or for staying informed about world news. Therefore, to promote the growth of this new way of conducting commerce, it is undoubtedly a good approach to encourage investments in a new market. As we come to understand, the value of cryptocurrencies is not regulated like traditional currency, and much of their value depends on the demand from investors, which can be used to create a "bubble" in their value [3].

Many of the top cryptocurrency influencers are investors or creators themselves. One example is Vitalik Buterin, who is the creator of Ethereum [4], which is one of the most well-known cryptocurrencies worldwide. We can see that Vitalik has a deep understanding of cryptocurrencies and blockchain, making him an influencer in this field. With his comments on social media, he can contribute to or harm the value of his cryptocurrency in the market. Due to the free use of social media and the ability for anyone to express their opinions on any topic, it becomes difficult to determine who truly has influence and the impact they can have on their community. Financial

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

espinosagonzalezdaniel@gmail.com (D. Y. Espinosa); sidorov@cic.ipn.mx (G. Sidorov)
http://www.cic.ipn.mx/~sidorov/ (G. Sidorov)

bubbles, mainly driven by speculation, can rise or fall based on societal speculation on social media [3].

On this occasion, PAN has decided to launch the task "Profiling Cryptocurrency Influencers with Few-shot Learning" where it proposes a series of influencer profiling and other classifications with limited training data [5] [1]. To solve this task, we decided to use the BERT model with different text configurations, which we will detail later on.

2. Corpus

To solve this task, PAN introduced a corpus that is divided into three subtasks to classify:

1. For the first subtask, there are 160 users with the following classifications: non-influencer, nano, micro, mega, and macro.
2. In the second subtask, we have 320 users with the following classifications: technical information, price update, trading matters, gaming, and other.
3. For the last subtask, we have a total of 256 users with the following classifications: subjective opinion, financial information, advertising, and announcement.

The complexity of this task lies in the number of samples. In this case, we have one tweet to perform the classification of each user.

3. Methodology

Regarding Twitter, in previous tasks, we had been using different classifiers and methods to profile users for different tasks. For example, in PAN 2019 [6], we used an N-gram configuration with various classifiers such as RandomForest, SVM, and LinearSVM, but we had much more data for the classifications[7]. Due to this, we decided to use BERT [8], with a small amount of training data, methods that incorporate "Attention" are assumed to be the most suitable for finding a solution[9].

3.1. Pre-processing steps

As in every text-related task, we always recommend using a preprocessing layer before directly using the models. The use of preprocessing on the data helps transform, clean, and prepare it for analysis or modeling. This helps improve the data quality and facilitates its use in algorithms and models.

The next configuration is using for the three subtasks:

Lowercase Were transform all tweets into lowercase to standardize the texts for the model.

Links were changed by the label 'link' within the data.

User Mentions are changed by the 'usermention' label that the tweets contain.

Hashtags the 'hashtag' label was modified and placed.

Emojis The emojis were modified and the label 'emoji' was placed.

Other symbols The symbols that are not registered within the ASCI reference standard are eliminated within the data set.

4. Experiments

In the past task, we used N-gram structures, used N-grams of words and N-grams of characters, mainly for tasks that involved social networks, since they showed outstanding results with this type of arrangement.The results of these experiment are in the **Table 1** and **Table 2**. All the results obtained in this work were tested with **Accuracy** except **Table 5** where show the results with F1 because this is the metric which PAN evaluates the results[1].

Since the results with the combination of Ngrams did not have the expected precision, it was thought to implement another model to carry out the profiling.

In PAN 2022[10], Wang Bin [11] used BERTweet [12] obtaining a result close to 95% accurate with this methodology, so he experimented with it, **Table 3** shows the results.

Although the results with BERTweet were better, we decided to test with the BERT configuration as we wanted to compare a similar model across all three subtasks. We used a batch size configuration of 16. Previously, we used 32, but with a batch size of 32 and 8 epochs, it required more than 22 GB of GPU memory.

All our experiments were conducted using GPU to accelerate processing and enable parallelization of the processes.

Table 1
Results of combinations **N-grams character** accuracy

N-grams character	subtask1	subtask2	subtask3
2-3-4-5-9	61.22	63.12	61.98
3-4-5-9	62.45	62.91	62.38
5-7-8-9	64.37	65.74	64.40
8-9	71.21	71.95	72.06

Table 2
Results of combinations **N-grams words** accuracy

N-grams words	subtask1	subtask2	subtask3
2-3-4-5-9	71.87	72.90	71.99
3-4-5-9	81.25	83.02	82.76
5-7-8-9	80.30	81.07	82.28
8-9	82.71	83.99	83.54

Table 3
Results of **BERTweet** accuracy with and without preprocessing layer

task	without preprocessing layer	with preprocessing layer
subtask 1	86.21	86.98
subtask 2	87.12	88.76
subtask 3	90.14	92.33

Table 4
Results of **BERT** accuracy with and without preprocessing layer

task	without preprocessing layer	with preprocessing layer
subtask 1	89.34	**92.34**
subtask 2	93.21	**95.44**
subtask 3	95.88	**96.94**

Table 5
Results of **BERT** F1 with and without preprocessing layer

task	without preprocessing layer	with preprocessing layer
subtask 1	79.02	**86.77**
subtask 2	81.33	**83.44**
subtask 3	81.54	**85.01**

Table 6
Results of **DistilBERT, TinyBERT and RoBERTa** accuracy with preprocessing layer

task	DistilBERT	TinyBERT	RoBERTa
subtask 1	83.29	75.21	**93.14**
subtask 2	85.12	76.84	**95.61**
subtask 3	85.98	77.23	**97.03**

Furthermore, by reducing the batch size, we were able to add an extra epoch, significantly reducing GPU consumption while maintaining precision at a very similar level. Both the BERTweet and BERT results were obtained using **9 epochs**, and with this configuration, we achieved the best results without overfitting the models. The results we chose to participate with were from the BERT model, and the results for all three subtasks are shown in **Table 4**.

Regarding the results using BertTweet, which were not as expected, we tried other BERT configurations to see if we could improve the results. Therefore, we used three more variations: DistilBERT[13], RoBERTa Liu et al. [14], and TinyBERT[15].

As shown in **Table 6**, the results with DistilBERT and TinyBERT were very close but not the best for the task. On the other hand, with RoBERTa, the results improved compared to BERT, but they were not outstanding either. The issue with RoBERTa was the time and resource requirements. With BERT and the mentioned hyperparameters, the training took around 45 minutes. In contrast, with RoBERTa, it took approximately 4 hours and 40 minutes. These tests were performed using a 13 GB GPU. Consequently, we decided not to use RoBERTa because we did not consider its results to be significantly better than BERT.

5. Conclusions

Given the mechanism of BERT, it is a powerful option for tweet classification due to its ability to understand context and utilize attention, which can be highly beneficial when there is limited training data. These combined advantages can significantly enhance the accuracy and classification capability of tweets. Despite the challenge of training with limited data, it is interesting and useful that models work with these characteristics since data collection can be costly and, in some cases, difficult. This makes machine learning more straightforward and accessible, allowing for faster adaptation to the real world.

An outstanding observation from these experiments was the use of BERTweet [12] and BERT[8]. Although both models are based on the Transformer architecture, they were trained on different datasets, highlighting their distinct approaches. While the accuracy of BERTweet was not bad, we definitely preferred using BERT as it demonstrated better precision with limited data. These findings are highly interesting, and in our opinion, they should be utilized in other experiments.

The use of social media is filled with opinions, highlighting the importance of learning who shares the truth. In the case of cryptocurrency influencers, they can be a reliable source of information, but the challenge lies in knowing who is truly an expert or a trustworthy person on the internet. These types of experiments promote the use of truth on social media, although when it comes to investments, it is best to seek advice and assistance from a professional in the field before investing or making decisions on these matters.

References

[1] M. Chinea-Rios, I. Borrego-Obrador, M. Franco-Salvador, F. Rangel, P. Rosso, Profiling Cryptocurrency Influencers with Few shot Learning at PAN 2023, in: CLEF 2022 Labs and Workshops, Notebook Papers, 2023.

[2] G. L. W. Y. Kuo Chuen, David LEE, Cryptocurrency: A new investment opportunity?, 2018. URL: https://ink.library.smu.edu.sg/lkcsb_research/5784. doi:10.3905/jai.2018.20.3.016.

[3] R. Sawhney, S. Agarwal, V. Mittal, P. Rosso, V. Nanda, S. Chava, Cryptocurrency bubble detection: A new stock market dataset, financial task & hyperbolic models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5531–5545. URL: https://aclanthology.org/2022.naacl-main.405. doi:10.18653/v1/2022.naacl-main.405.

[4] V. Buterin, N. Schneider, Proof of Stake: The Making of Ethereum and the Philosophy of Blockchains, Seven Stories Press, 2022. URL: https://books.google.com.mx/books?id=hWpVEAAAQBAJ.

[5] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2023: Authorship

Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, A. G. Stefanos Vrochidis, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, Springer, 2023.

[6] F. Rangel, P. Rosso, CLEF 2019 Labs and Workshops, Notebook Papers, in: C. L., F. N., M. H, L. D. (Eds.), Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling, CEUR Workshop Proceedings, 2019.

[7] D. Espinosa, H. Gómez-Adorno, G. Sidorov, Bots and Gender Profiling using Character Bigrams, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[10] R. O. Bueno, B. Chulvi, F. Rangel, P. Rosso, E. Fersini, Profiling irony and stereotype spreaders on twitter (irostereo). overview for pan at clef 2022, ????, pp. 2314–2343. URL: http://ceur-ws.org/Vol-3180/#paper-185.

[11] W. Bin, N. Hui, Notebook for pan at clef 2022:profiling irony and stereotype spreaders on twitter, ???? URL: https://ceur-ws.org/Vol-3180/paper-225.pdf.

[12] D. Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14. URL: https://aclanthology.org/2020.emnlp-demos.2. doi:10.18653/v1/2020.emnlp-demos.2.

[13] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. arXiv:1910.01108.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[15] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, Tinybert: Distilling bert for natural language understanding, 2020. arXiv:1909.10351.