Text Enrichment with Japanese Language to Profile Cryptocurrency Influencers

Notebook for PAN at CLEF 2023

Francesco Lomonaco[1], Marco Siino[2,3] and Maurizio Tesconi[2]

[1]*Università degli Studi di Milano Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione, Milano, 20126, Italy*

[2]*Istituto di Informatica e Telematica, CNR, Pisa, 56127, Italy*

[3]*Università degli Studi di Palermo, Dipartimento di Ingegneria, Palermo, 90128, Italy*

Abstract

From a few-shot learning perspective, we propose a strategy to enrich the latent semantic of the text provided in the dataset provided for the Profiling Cryptocurrency Influencers with Few-shot Learning, the task hosted at PAN@CLEF2023. Our approach is based on data augmentation using the backtranslation forth and back to and from Japanese language. We translate samples in the original training dataset to a target language (i.e. Japanese). Then we translate it back to English. The original sample and the backtranslated one are then merged. Then we fine-tuned two state-of-the-art Transformer models on this augmented version of the training dataset. We evaluate the performance of the two fine-tuned models using the Macro and Micro F1 accordingly to the official metric used for the task. After the fine-tuning phase, ELECTRA and XLNet obtained a Macro F1 of 0.7694 and 0.7872 respectively on the original training set. Our best submission obtained a Macro F1 equal to 0.3851 on the official test set provided.

Keywords

cryptocurrency influencers, data augmentation, author profiling, text classification, Twitter, text enrichment, japanese

1. Introduction

The task proposed at PAN@CLEF2023 [1] was about Profiling Cryptocurrency Influencers with Few-shot Learning on Twitter [2]. The task was to profile cryptocurrency influencers in social media, from a low-resource perspective. The task organizers proposed three multilabel classification subtasks: 1) Low-resource influencer profiling, 2) Low-resource influencer interest identification, 3) Low-resource influencer intent identification. All the subtasks are multilabel classification tasks requiring strategies based on few-shot learning considering the size of the three dataset provided. In fact, the authors in the three corresponding datasets were, respectively, 32, 64 and 64. Furthermore, the number of tweets available for each author was never above ten. With such a low-resource perspective some form of transfer-learning was definitely required.

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

marco.siino@unipa.it (M. Siino)

https://github.com/marco-siino (M. Siino)

0000-0002-2295-1443 (F. Lomonaco); 0000-0002-4453-5352 (M. Siino); 0000-0001-8228-7807 (M. Tesconi)

CEUR Workshop Proceedings (CEUR-WS.org)

The rest of this work is organized as follows. In Section 2 we briefly present related work about text classification, along with a brief discussion on some of the architecture proposed in the previous editions of PAN. In Section 3 we describe our framework, including training and inferencing stages. In Section 4 we discuss the experimental evaluation of our framework, reporting the results obtained. In Section 5 we propose future works and conclude the paper.

2. Related Work

With our submission we extend the one conducted in [3]. Considering that the proposed task hosted at PAN@CLEF2023 consists of a few-shot learning one, we used the mentioned work as a starting point. In this work the authors propose a data augmentation technique based on backtranslation to augment samples in the dataset. The authors made use of the Italian language for the backtranslation proving the effectiveness of this strategy when compared to a non-augmented version of the training dataset. However, for our submission we wanted to investigate the impact of a morphologically different language as Japanese. Because, just like English, Japanese has no gender and there is no differentiation between plural and singular but, compared to English, in Japanese word order is normally subject–object–verb.

For the above-mentioned reasons, our participation at PAN@CLEF2023 [4, 5] is based on a text data augmentation strategy, plus a further stage which consists in the use of two state-of-the-art Transformers (namely, XLNet [6] and ELECTRA [7]). We looked into the top-performing models that were taking part in the shared tasks presented by PAN to construct the method we propose. We examined the outcomes of the top team in the author profiling challenge held during PAN@CLEF 2021, where the proposed architecture consisted of a shallow CNN presented in [8, 9]. We also looked at the winning model at PAN@CLEF2022 where the authors won the competition because of a soft voting ensemble technique that combines BERTweet models with various loss functions and a BERT feature-based CNN model. In the 2020 author profiling edition [10], based on their most recent 100 tweets, the aim was to identify authors who were likely to share false information. The authors of [11, 12] were the winners of the proposed task. On the supplied test set, their models achieved a 0.77 overall accuracy. The winning strategies are based on an ensemble of various machine learning models, an SVM, n-grams, and other techniques. Other ensemble models have been proposed at the following year task hosted at PAN about irony and stereotype spreaders detection [13, 14].

In [15] SVM, Naive Bayes, Logistic Regression, and Recurrent Neural Networks (RNN) are just a few of the popular machine learning algorithms that the authors compare. SVM and Naive Bayes perform better than other algorithms on the dataset used, according to experimental results. In addition to the RNN, authors do not mention evaluation of a CNN or models based on deep learning. In another relevant comparative study [16], on three separate datasets, scholars assess seven machine learning models. The models employed are based on Random Forest, SVM, Gaussian Naive Bayes, AdaBoost, KNN, Multi-Layer Perceptron, and Gradient Boosting Algorithm. The Gradient Boosting Algorithm performs better than the other examined models in terms of accuracy and F1 score.

In [17], in order to profile spreaders of fake news, the authors feed a CNN using psycholinguistic and linguistic characteristics. The outcomes of their experiments demonstrate that their suggested strategy is successful in identifying users sharing misinformation. On a dataset created especially for their goal, the authors compare their findings. The performance of deep models is not extensively examined, and BERT is the only Transformer that has been tested. The proposed architecture is also tested in [18] (where the PAN@CLEF2020 dataset is used). Accordingly to the paper, the model is able to reach 0.52 and 0.51 of accuracy on the English and on the Spanish dataset respectively. Within the study [18], the authors suggest a novel approach that performs better for both languages than the two PAN@CLEF2020 winning models by utilizing personality data and visual elements.

In the work conducted in [19], for the purpose of classifying sentiments, authors suggest using CNN. The authors demonstrate that consecutive convolutional layers are useful for classifying lengthy texts.

Additionally, we examined numerous contemporary methods for text classification problems. The usage of Explainable Artificial Intelligence (XAI) techniques rather than black box-based strategies has increased significantly, and this is noteworthy to note. Several of these techniques have graph-based foundations and have practical applications in social networking [20], text classification [21], computer vision [22] and traffic prediction [23].

With regard to cryptocurrency, authors in [24] create a series of sequence-to-sequence hyperbolic models based on the power-law dynamics of cryptocurrencies and user activity on social media that are appropriate for bubble detection identification challenges. An interesting study from an NLP perspective is the one presented in [25]. To understand the relationships between cryptocurrency values and social media, the authors analyze what happened in social media starting in June 2019 using a combination of statistical models and NLP techniques, concentrating on the rise of the Ethereum and Bitcoin prices.

We chose to use XLNet and ELECTRA as classifiers in light of the results obtained in a similar multi-label text classification task [26] and presuming, as discussed in [27, 28], that deep AI models are actually capable of outperforming traditional techniques used in the field of NLP.

3. The Proposed Approach

The proposed framework is evaluated through a three-stage empirical experiment. First, the baseline of author profiling models are established using datasets without our augmentation modules. The second stage involves generating augmented data using backtranslation to and from Japanese. Backtranslating to and from a morphologically so different can produce very interesting results in terms of latent semantic that is made explicit thanks to the translation [29]. After the backtranslation of the original sample, the backtranslated version is merged to the original one. Then the augmented data is used to train the XLNet and ELECTRA to compare the performances with or without the backtranslation module. We use ELECTRA for our first submission and XLNet for the second one. This choice was dictated by the results presented in [30] where the authors report the remarkable performance of ELECTRA for one-shot-learning tasks. Similarly, XLNet has shown interesting results on another few-shot-learning-oriented task as reported in [26]. We did not use other larger Transformer models (as BERT, T5 or RoBERTa) to better asses the impact of the Japanese language. In our setting, each sample is a user's set of tweets, and we hypothesise that semantically enriching the user's tweets using

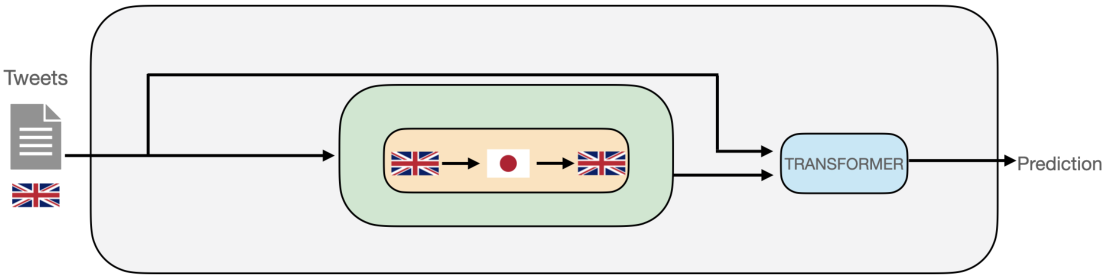

Figure 1: The overall architecture of the proposed framework. For our submission, the backtranslation module translates the original source text into Japanese and then back into English. Then XLNet and ELECTRA are separately fine-tuned on the augmented version of the training set.

our proposed modules can improve performance. By augmenting each sample with one or multiple translations, we aim to increase the diversity and informativeness of the data and improve the representation of the input, ultimately leading to better classification performance of different NLP models. Our results outperform the not-augmented baseline, showing that the expansion of samples with multiple languages using backtranslation leads to improved performances in author profiling tasks. Thanks to the backtranslation module our framework is able to outperform the results obtained without expanding the samples.

No preprocessing is applied to the source text in the training datasets. In Figure 1 we show the frameworks we used for our two submissions at the subtask 1. In the first submission we used the augmented training set for fine-tuning and for inferencing using an ELECTRA Transformer, for the second submission we used an XLNet instead.

Our model is trained on the augmented versions of the datasets. For both the submissions we fine-tuned ELECTRA and XLNet for 30 epochs on the augmented dataset. To perform the translation forth and back to and from Japanese we used the Google Translate API[1]. After the training phase, we used the fine-tuned Transformers to predict on the unlabeled test set provided by the task organizers.

[1] https://pypi.org/project/googletrans/

4. Experimental Evaluation

4.1. Experimental Setup

Our training and inferencing notebooks, developed in TensorFlow and using the *Simple Transformers*[2] library, are publicly available as a Jupyter Notebook on GitHub[3]. For the training and for the inferencing phases we made use of ELECTRA and XLNet [6]. According to what stated in [7], ELECTRA suggests to replace certain tokens with possible replacements taken from a small generator network, instead of masking input like in BERT. Then, a discriminative model is trained to predict whether each token in the corrupted input was replaced by a generator sample or not, as opposed to developing a model that predicts the original identities of the corrupted tokens. Along with a graph neural network, ELECTRA can also be employed as an embedding layer as in [21]. In our experiments, the original version of ELECTRA, presented in [7], was used. XLNet was developed by optimizing the predicted likelihood across all combinations of the factorization order to enable learning bidirectional contexts. XLNet surpasses BERT, frequently by a significant margin, on a number of tasks, including question answering, sentiment analysis, document ranking and natural language inference. For our study we used the pre-trained XLNet using the zero-shot cross lingual transfer discussed in [31]. In both cases we used a batch size of 1. We fine-tuned both models for 30 epochs. No improvements are obtained in fine-tuning for more epochs.

4.2. The Dataset

The PAN organizers' dataset includes a list of Twitter authors and a variable number of corresponding tweets. For each author in the training set the labels are also provided. All the details are reported on the official task website[4]. With regard to the three proposed subtask, they were, namely: 1) Low-resource influencer profiling, 2) Low-resource influencer interest identification, 3) Low-resource influencer intent identification. For the first subtask the organizers provided an English dataset with 32 users per label with a maximum of 10 English tweets each. The five labels available were: (1) null, (2) nano, (3) micro, (4) macro, (5) mega depending on the type of influencer the author was. For the second subtask were provided 64 users per label with 1 English tweets each. The five labels available in this case were: (1) technical information, (2) price update, (3) trading matters, (4) gaming, (5) other. Finally, for the third subtask, 64 users per label with 1 English tweets each were available and the classes available for predictions were: (1) subjective opinion, (2) financial information, (3) advertising, (4) announcement. In this paper we discuss the framework we used to participating at the first subtask (i.e., low-resource influencer profiling).

4.3. Results

The Macro F1 is the adopted metric for the author profiling task at PAN@CLEF2023. This metric, along with accuracy, is the same used in the rest of this section and defined in (1).

[2] https://simpleTransformers.ai/about/

[3] https://github.com/marco-siino/PAN-CRYPTO-2023

[4] https://pan.webis.de/clef23/pan23-web/author-profiling.html

Table 1
Results achieved in terms of F1 per each class at the end of the fine-tuning of the two Transformers on the augmented training set. The evaluation is performed using precision and recall on the non-augmented version of the training set.

F1 results per each class on the original non-augmented training set					
	Macro	Nano	No	Mega	Micro
ELECTRA	0.8108	0.8064	0.7692	0.6800	0.7805
XLNet	0.7631	0.7451	0.8750	0.7857	0.7671

Table 2
Macro results achieved by our framework at the end of the fine-tuning on the augmented dataset using Japanese. The results reported are obtained evaluating the fine-tuned Transformers on the original training set.

Results on the original training set		
	Macro F1	Acc
ELECTRA	0.7694	0.7750
XLNet	0.7872	0.7875

$$MacroF1 = \frac{sum(F1scores)}{\#classes} \tag{1}$$

In Table 1 we report the results using the F1 related to the five classes. The F1 is calculated, using the official evaluator, for all the classes available and using the original non-augmented version of the training set provided by the task organizers. The Python code of the evaluator is available on GitHub[5].

Finally, in Table 2, we report the results with the metrics related to all the classes (i.e. Macro F1, accuracy), evaluating the results for the original non-augmented version of the training set.

Although the Macro F1 and the accuracy prove that XLNet fine-tuned on the Japanese back-translated version of the dataset outperforms ELECTRA, as can be seen from Table 2 for three out of five classes the Micro F1 is higher using ELECTRA. However, a further investigation on the effect of the backtranslation on the original samples could eventually lead to an explanation of these differences among the classes. Finally, according to the final ranking[6] it is worth mentioning that on the unlabelled test set provided, our best submission reached a Macro F1 equal to 0.3851.

5. Conclusion and Future Works

In this paper we have described our submitted model for our participation at the author profiling task hosted at PAN@CLEF 2023. It consists of a backtranslation layer followed by an expansion module to expand every sample in the dataset. These augmented versions of the samples are then provided to ELECTRA and XLNet both for training and inference phase.

[5] https://github.com/pan-webis-de/pan-code/tree/master/clef23/profiling-cryptocurrency-influencers
[6] https://pan.webis.de/clef23/pan23-web/author-profiling.html

We plan to assess performance using other backtranslation methods and additional languages in future works. Improved performance on the proposed classification task may even result from conducting an error analysis on authors who were incorrectly classified. The size of the dataset made available allowed for the application of certain other data augmentation techniques. Before the training and testing phases of our model, some research into the content of each tweet may help steer the construction of the model by applying various strategies to remove noise (i.e., irrelevant characteristics) from input samples. We found that enriching samples with their respective backtranslations can lead to performance improvements.

It would be interesting to further explore the matter also using additional datasets relevant to those used for author profiling tasks. Eventually, it could also be interesting to assess the effect of using different languages in the backtranslation module.

Acknowledgments

We would like to thank anonymous reviewers for their comments and suggestions that have helped to improve the presentation of the paper.

CRediT Authorship Contribution Statement

Francesco Lomonaco: Investigation, Resources, Software. **Marco Siino:** Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - Original draft, Writing - review & editing. **Maurizio Tesconi:** Writing - review & editing.

References

[1] 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.

[2] 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 2023 Labs and Workshops, Notebook Papers, 2023.

[3] S. Mangione, M. Siino, G. Garbo, Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network, in: CEUR Workshop Proceedings, volume 3180, CEUR, 2022, pp. 2585–2593.

[4] M. Siino, M. Tesconi, I. Tinnirello, Profiling cryptocurrency influencers with few-shot learning using data augmentation and electra, in: CLEF 2023 Labs and Workshops, Notebook Papers, 2023.

[5] M. Siino, I. Tinnirello, Xlnet on augmented dataset to profile cryptocurrency influencers, in: CLEF 2023 Labs and Workshops, Notebook Papers, 2023.

[6] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[7] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[8] M. Siino, E. Di Nuovo, I. Tinnirello, M. La Cascia, Detection of hate speech spreaders using convolutional neural networks, in: PAN 2021 Profiling Hate Speech Spreaders on Twitter@ CLEF, volume 2936, CEUR, 2021, pp. 2126–2136.

[9] M. Siino, E. Di Nuovo, I. Tinnirello, M. La Cascia, Fake news spreaders detection: Sometimes attention is not all you need, Information 13 (2022) 426.

[10] F. Rangel, A. Giachanou, B. H. H. Ghanem, P. Rosso, Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, in: CEUR Workshop Proceedings, volume 2696, Sun SITE Central Europe, 2020, pp. 1–18.

[11] J. Pizarro, Using n-grams to detect fake news spreaders on twitter, in: CLEF, 2020, p. 1.

[12] J. Buda, F. Bolonyai, An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter, in: CLEF, 2020, p. 1.

[13] D. Croce, D. Garlisi, M. Siino, An svm ensamble approach to detect irony and stereotype spreaders on twitter, in: CEUR Workshop Proceedings, volume 3180, CEUR, 2022, pp. 2426–2432.

[14] M. Siino, I. Tinnirello, M. La Cascia, T100: A modern classic ensemble to profile irony and stereotype spreaders, in: CEUR Workshop Proceedings, volume 3180, CEUR, 2022, pp. 2666–2674.

[15] E. M. Mahir, S. Akhter, M. R. Huq, et al., Detecting fake news using machine learning and deep learning algorithms, in: 2019 7th International Conference on Smart Computing & Communications (ICSCC), IEEE, 2019, pp. 1–5.

[16] A. P. S. Bali, M. Fernandes, S. Choubey, M. Goel, Comparative performance of machine learning algorithms for fake news detection, in: International conference on advances in computing and data sciences, Springer, 2019, pp. 420–430.

[17] A. Giachanou, B. Ghanem, E. A. Ríssola, P. Rosso, F. Crestani, D. Oberski, The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers, Data & Knowledge Engineering 138 (2022) 101960.

[18] R. Cervero, P. Rosso, G. Pasi, Profiling Fake News Spreaders: Personality and Visual Information Matter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2021, pp. 355–363.

[19] H. Kim, Y.-S. Jeong, Sentiment classification using convolutional neural networks, Applied Sciences 9 (2019) 2347.

[20] M. Siino, M. La Cascia, I. Tinnirello, Whosnext: Recommending twitter users to follow using a spreading activation network based approach, in: 2020 International Conference on Data Mining Workshops (ICDMW), IEEE, 2020, pp. 62–70.

[21] F. Lomonaco, G. Donabauer, M. Siino, Courage at checkthat! 2022: Harmful tweet detection using graph neural networks and electra, in: Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022, pp. 573–583.

[22] P. Pradhyumna, G. Shreya, et al., Graph neural network (gnn) in image and video understanding using deep learning for computer vision applications, in: 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 2021, pp. 1183–1189.

[23] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, arXiv preprint arXiv:1707.01926 (2017).

[24] R. Sawhney, S. Agarwal, V. Mittal, P. Rosso, V. Nanda, S. Chava, Cryptocurrency bubble detection: a new stock market dataset, financial task & hyperbolic models, arXiv preprint arXiv:2206.06320 (2022).

[25] M. Ortu, S. Vacca, G. Destefanis, C. Conversano, Cryptocurrency ecosystems and social media environments: An empirical analysis through hawkes' models and natural language processing, Machine Learning with Applications 7 (2022) 100229.

[26] M. Siino, M. La Cascia, I. Tinnirello, McRock at SemEval-2022 task 4: Patronizing and condescending language detection using multi-channel CNN, hybrid LSTM, DistilBERT and XLNet, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 409–417. URL: https://aclanthology.org/2022.semeval-1.55. doi:10.18653/v1/2022.semeval-1.55.

[27] H. Wu, Y. Liu, J. Wang, Review of text classification methods on deep learning, CMC-Computers, Materials & Continua 63 (2020) 1309–1321.

[28] S. Hashida, K. Tamura, T. Sakai, Classifying tweets using convolutional neural networks with multi-channel distributed representation, IAENG International Journal of Computer Science 46 (2019) 68–75.

[29] M. Seligman, The evolving treatment of semantics in machine translation, Adv. Empir. Transl. Stud. Dev. Transl. Resour. Technol. (2019) 53.

[30] S. Ni, H.-Y. Kao, Electra is a zero-shot learner, too, arXiv preprint arXiv:2207.08141 (2022).

[31] G. Chen, S. Ma, Y. Chen, L. Dong, D. Zhang, J. Pan, W. Wang, F. Wei, Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 15–26.

A. Online Resources

The source code of our model is available via

- GitHub