XLNet with Data Augmentation to Profile Cryptocurrency Influencers

Notebook for PAN at CLEF 2023

Marco Siino[1,2], Ilenia Tinnirello[1]

[1]Department of Engineering, University of Palermo, Palermo, 90133, Italy
[2]Cyber Intelligence Lab, Institute of Informatics and Telematics, National Research Council, Pisa, 56127, Italy

Abstract

In this work we propose an application of XLNet to address the task hosted at PAN@CLEF2023 related to Profiling Cryptocurrency Influencers with Few-shot Learning. For our proposed approach we made use of XLNet fine-tuned on an augmented version of the training set provided for the competition. Given the few-shot learning perspective of the task we found useful to employ a data augmentation strategy similar to one proposed in a previous edition of a PAN task. The augmentation is performed augmenting each sample in the training dataset with its corresponding backtranslated version to a target language. The target languages we used for our two submissions were German and Italian. After the fine-tuning of the XLNet we predict the labels for the unlabeled test set. After fine-tuning the XLNet model we evaluated it on the original non-augmented training set. We evaluated all the F1 with regards to each label, and then we reported the Macro F1 across all the labels provided. Our results prove that on the original training set our approach can obtain a maximum Macro F1 of 0.6937 and a maximum accuracy of 0.6893.

Keywords

cryptocurrency influencers, few-shot learning, author profiling, text classification, Twitter, text data augmentation, xlnet

1. Introduction

The author profiling challenge proposed at PAN@CLEF2023 [1] was about Cryptocurrency Influencers with Few-shot Learning on Twitter [2]. Identifying cryptocurrency influencers on social media was the assignment from a low-resource perspective. The task organizers proposed three multi-label classification subtasks. They were, namely: 1) Low-resource influencer profiling, 2) Low-resource influencer interest identification, 3) Low-resource influencer intent identification. Regarding 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,

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
marco.siino@unipa.it (M. Siino); ilenia.tinnirello@unipa.it (I. Tinnirello)
https://github.com/marco-siino (M. Siino)
0000-0002-4453-5352 (M. Siino); 0000-0002-1305-0248 (I. Tinnirello)

(4) gaming, (5) other. Finally, for the third subtask, 64 users per label with 1 English tweets each were available and 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 participate in the first subtask (i.e., low-resource influencer profiling).

The rest of this paper is organized as follows. In Section 2 we report some related works on text classification methods, along with a brief discussion on some of the architecture proposed in the previous editions of PAN. In Section 3 we introduce our proposed approach. In Section 4 we detail the experimental setup and evaluation, concluding the section discussing the results obtained. In Section 5 we discuss future works concluding the paper.

2. Related Work

The power-law dynamics of cryptocurrencies and user activity on social media are used by the authors of [3] to construct a variety of sequence-to-sequence hyperbolic models that are appropriate for bubble detection identification problems. From the perspective of NLP, the study mentioned in [4] is intriguing. In order to better understand the relationships between cryptocurrency values and social media, the authors employ a combination of statistical models and NLP techniques to look at what happened in social media starting in June 2019 with a focus on the rise of the Ethereum and Bitcoin prices.

The proposed task at PAN@CLEF2023 is presented in [5]. Furthermore, to develope our proposed approach we looked at the top performing teams that participated in the previous shared tasks organized by PAN. We noticed the results of the winning team at the 2021 edition, where the winning model consisted of a shallow CNN presented and discussed in [6, 7]. We also looked at the winning model at PAN@CLEF2022 where the authors won the contest thanks to a soft voting ensemble technique that combines BERTweet models with various loss functions and a BERT feature-based CNN model. The aim of the author profiling task in a previous edition [8] was to identify writers who were likely to propagate false information based on their most recent 100 tweets. The winning teams were [9] and [10]. On the given test set, their models accuracy was 0.77. The winning strategies were based on n-grams, an SVM, and an ensemble of other machine learning models. Other ensemble models have been proposed for other shared tasks at PAN in the following years [11, 12].

We also examined a number of contemporary models for text categorization problems. It is important to note that Explainable Artificial Intelligence (XAI) techniques are increasingly being used in place of black box-based strategies. Several of these techniques based on graphs are applied in actual applications like text classification [13], traffic prediction [14], computer vision [15] and social networking [16]. Authors in [17] compare SVM, Naive Bayes, Logistic Regression, and Recurrent Neural Networks (RNN) as well as other popular machine learning methods. Experimental results demonstrate that SVM and Naive Bayes outperform other approaches on the dataset employed. In addition to the RNN, they do not report the evaluation of CNN or deep learning-based models. In another relevant comparative study [18], on three separate datasets, scholars assess seven machine learning methods. Random Forest, SVM, Gaussian Naive Bayes, AdaBoost, KNN, Multi-Layer Perceptron, and Gradient Boosting Algorithm are among the models that were utilized. The Gradient Boosting Algorithm surpasses the other examined models in terms of accuracy and F1 score. There are not any additional deep model experiments in this work, though.

In [19] the task of automatically detecting fake news spreaders of COVID-19 news is addressed by the authors by extending the CoAID dataset[20]. The authors offer a stacked Transformer-based neural network that combines a deep learning model with the ability of Transformer to compute language embeddings.

In [21], the authors profile fake news spreaders using psycholinguistic and linguistic characteristics as input to CNN. The outcomes of their experiments demonstrate how the suggested model categorizes users as fake news spreaders. The dataset used for the authors' comparison was created expressly for their goal. However, only BERT was tested as transformer, and little is known about the performance of deep models. Their model has also been model evaluated on the PAN2020 dataset in [22] reporting results. On the English and Spanish datasets, the tested model achieves a binary accuracy of 0.52 and 0.51, respectively. In the same work [22], the authors suggest a novel model that outperforms the two winning models at PAN@CLEF2020 on both languages by utilizing personality data and visual features.

In the work conducted in [23], for the purpose of sentiment classification, scholars suggest using a CNN. The authors demonstrate that using consecutive convolutional layers is efficient for categorizing lengthy texts through tests with three well-known datasets.

Finally, the survey in [24] gives a succinct rundown of various text classification algorithms. This overview discusses several ways for extracting text features, dimensionality reduction, existing algorithms and methods, and evaluation strategies.

Given the performances shown in another international multi-label text classification challenge [25] and, as discussed in [26, 27], presuming that natural language processing conventional methods can truly be outperformed by deep AI models, we decided to employ a Transformer based architecture (namely, XLNet [28]). Considering that the proposed task hosted at PAN@CLEF2023 consists on few-shot learning we also evaluated the augmented technique discussed in [29]. In this work the authors propose a data augmentation technique based on backtraslation to augment samples in the dataset.

3. The Proposed Approach

The framework we propose is evaluated through a three-stage empirical experiment. First, baseline of author profiling models are established using the provided dataset without our augmentation modules. The second stage involves generating augmented data using backtranslation from English to a target language. The backtranslated sample is then concatenated to the original one. In the final stage, the augmented data are used to train XLNet [28] and to compare the performances between different target languages. In our setting, each sample is a user's set of tweets, and we hypothesise that semantically enriching the user's tweets using our proposed modules can improve performance. By augmenting each sample with a backtranslation, 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 show that the expansion of samples with multiple languages using backtranslation significantly impact on performances in author profiling tasks [30, 31]. 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. To select the two target languages (i.e., German and Italian) we considered the studies in [29, 32]. In the first submission we augmented the training set backtranslating from German and in the second submission, we backtranslated from Italian as done in [29] where as a last stage classifier the authors did not use a Transformer but a shallow CNN instead.

The training of our model is performed on the augmented versions of the datasets. For the first submission we fine-tuned the XLNet for 30 epochs on the dataset augmented using the backtranslation tecnique with the German language. For the second one we used the Italian as a target language. In both cases we backtranslated the samples using the Google Translate API[1]. After the training phase, we used the fine-tuned XLNet to predict on the unlabeled test set provided by the task organizers.

4. Experimental Evaluation

4.1. Experimental Setup

Our training and inferencing models, developed in TensorFlow and using *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 **XLNet**. To enable learning bidirectional contexts, XLNet was created by optimizing the expected likelihood over all possibilities of the factorization order. On a variety of tasks, such as question answering, sentiment analysis, document ranking, and natural language inference, XLNet outperforms BERT, frequently by a wide margin. In our work, we used the zero-shot cross-lingual transfer-trained XLNet [33] given the promising results discussed in [25] for a similar multi-label text classification task. In both submissions we used a batch size of 1. We fine-tuned XLNet for 30 epochs with early stopping for the tuned version that we used for the final predictions. No improvements are obtained in fine-tuning for more epochs.

4.2. The Dataset

A set of Twitter authors made up the dataset along with a variable number of corresponding tweets. For each author in the training set the labels are also provided. The five labels available were: (1) null, (2) nano, (3) micro, (4) macro, (5) mega. The dataset consists of 32 users with a maximum of 10 English tweets each. Further details on the subtask are available at the official task website[4].

4.3. Results

For the author profiling task we discuss here, the Macro F1 is the official metric at PAN@CLEF2023. This metric, along with the Micro one, is the same used in the rest of this section.

[1]https://pypi.org/project/googletrans/
[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, per each label, achieved by our framework a the end of the fine-tuning using the two augmented versions of the original training set. The evaluation is performed on the non-augmented version of the training set.

F1 results per class on the non-augmented train set					
	Macro	Nano	No	Mega	Micro
German	0.4000	0.4390	0.7059	0.6914	0.6078
Italian	0.6122	0.7407	0.6667	0.7750	0.6517

Table 2
Results achieved by our framework in terms of Macro metrics. In this case the results are obtained using the original version of the training dataset for the evaluation.

Results on the original training set		
	Macro F1	Acc
German	0.5688	0.5937
Italian	0.6937	0.6893

In Table 1 we report the results using the F1 provided by the official evaluator available on GitHub[5] for all the classes available and using the original non-augmented version of the training set. Instead, in Table 2, we report the results using the Macro F1 and accuracy on the original non-augmented version of the training set.

Although the Macro F1 and the accuracy prove that XLNet fine-tuned on the Italian back-translated version of the dataset outperforms the German one, as can be seen from Table 1 for the "no-influencer" class the F1 is higher in the case of the German language. 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, on the Macro results an impressive difference exists between the two languages. The fine-tuned version of the XLNet using the augmented training set with Italian is able to outperform the German one by a large margin of more than 10% with respect to Macro F1 and around 9% in term of accuracy. According to the final ranking[6], our framework is able to obtain a Macro F1 of 0.3834 on the provided test set.

5. Conclusion and Future Works

In this paper we have presented our submitted model 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 XLNet both for the training and for the inference phase.

We intend to assess performance using different backtranslation techniques and other languages in future works. Even performing an error analysis on authors who were incorrectly classified can result in better scores on the suggested classification task. Increasing the model's

[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

complexity, perhaps by utilizing extra layers of Transformers, is another way to boost accuracy in author profiling. Given the size of the dataset that was provided, additional data augmentation techniques could possibly be used. Before the training and testing phases of our model, some research into the content of each tweet could influence the construction of the model in the use of some strategies to remove noise (i.e., not relevant features) from the input samples. We found that enriching samples with their respective backtranslations can lead to performance improvements.

As future works, it would be interesting to investigate this aspect also on other datasets related used for author profiling tasks. Furthermore, it could also be of interest to evaluate the impact of other languages used in the backtranslation module, although, as emerged from this study, the inclusion of a larger number of languages does not necessarily lead to an increase in the performance of the classification models.

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

Marco Siino: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - Original draft, Writing - review & editing. **Ilenia Tinnirello:** Conceptualization, Investigation, Methodology, 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] 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).

[4] 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.
[5] J. Bevendorff, M. Chinea-Ríos, M. Franco-Salvador, A. Heini, E. Körner, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, F. Rangel, et al., Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 518–526.
[6] 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.
[7] 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.
[8] 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.
[9] J. Pizarro, Using n-grams to detect fake news spreaders on twitter, in: CLEF, 2020, p. 1.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, arXiv preprint arXiv:1707.01926 (2017).
[15] 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.
[16] 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.
[17] 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.
[18] 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.

[19] S. Leonardi, G. Rizzo, M. Morisio, Automated classification of fake news spreaders to break the misinformation chain, Information 12 (2021) 248.

[20] L. Cui, D. Lee, Coaid: Covid-19 healthcare misinformation dataset, arXiv preprint arXiv:2006.00885 (2020).

[21] 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.

[22] 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.

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

[24] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, Information 10 (2019) 150.

[25] 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.

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

[27] 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.

[28] 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).

[29] 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.

[30] 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.

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

[32] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153.

[33] 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