Profiling Cryptocurrency Influencers with Sentence Transformers

Kavya Girish[1], Asha Hegde[2], Fazlourrahman Balouchzahi[3] and
Hosahalli Lakshmaiah Shashirekha[4]

Abstract

Few Shot Learning (FSL) is a supervised Machine Learning (ML) problem which deals with learning with few labeled samples. To address the challenges of FSL in terms of low-resource perspective, in this paper, we describe the models submitted to "Profiling Cryptocurrency Influencers with Few-shot Learning" - a shared task in PAN@CLEF 2023. The task has a focus on English Twitter posts for three subtasks: i) Subtask1 - Low-resource Influencer Profiling, ii) Subtask2 - Low-resource Influencer Interest Identification and iii) Subtask3 - Low-resource Influencer Intent Identification, with a very small set of labeled data. Two models: i) FSL-Word2Vec - Linear Support Vector Classifier (LinearSVC) trained with word embeddings extracted from Google's pre-trained Word2Vec and ii) FSL-ST - LinearSVC trained with sentence embeddings obtained from stsb-bert-base, are submitted to the shared task. FSL-Word2Vec models obtained macro F1 scores of 46.66 and 50.42 for Subtasks 2 and 3 respectively and FSL-ST model obtained a macro F1 score of 37.92 for Subtask1.

Keywords

Machine Learning, Cryptocurrency, Few Shot Learning, Word2Vec, Sentence Transformers

1. Introduction

Cryptocurrency is a digital currency created as an alternative form of currency using encryption algorithms. The transactions in cryptocurrency are verified and maintained by a decentralized system using cryptography, rather than by a centralized authority. Today, there are more than 22,789 different cryptocurrencies, with an estimated total value of 1 trillion dollars[1]. Some of the top cryptocurrencies with billions of dollars in the market include Bitcoin, Ethereum, Tether, Binance Coin, and Dogecoin.

Social media platforms are playing a prominent role in expanding the cryptocurrency communities to reach the general public. The rising ubiquity of speculative trading of cryptocurrencies over social media has led to sentiment driven "bubbles" [1, 2]. Further, comments/ posts in social media from highly influential personalities often cause a growing chain reaction leading to a short squeeze and the creation of a bubble. These personalities known as Cryptocurrency Influencers, act as the link between the cryptocurrency industry and general public. Vitalik

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

kavyamujk@gmail.com (K. Girish); hegdekasha@gmail.com (A. Hegde); fbalouchzahi2021@cic.ipn.mx (F. Balouchzahi); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)
https://sites.google.com/view/asha-hegde/home (A. Hegde); https://sites.google.com/view/fazlfrs/home (F. Balouchzahi); https://mangaloreuniversity.ac.in/shashirekha (H. L. Shashirekha)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://influencermarketinghub.com/top-crypto-influencers/

Buterin, Elon Musk, Andreas M, and S Anthony Pompliano are some of the most popular and leading Crypto Influencers. These Influencers can help the public and new crypto investors about the trending cryptocurrencies, comment on crypto news, and provide marketing services to crypto startups[2]. Cryptocurrency Influencers can be categorized based on: i) number of followers (Nano influencers - 1K–10K followers, Micro influencers - 10K–100K followers, Macro influencers - 100K–1M followers and Mega or celebrity influencers - 1M+ followers), ii) interest (technical information, price update, trading matters, gaming) and iii) intent (subjective opinion, financial information, advertising, announcement). People, especially the ones who want to know more information about cryptocurrencies or the ones who are interested in investing in cryptocurrencies can search for these Influencers to understand the nuances of crypto market. Hence, instead of searching for the Influencers, it will be very helpful if such Influencers' profile is automatically available to the users. This requirement leads to profiling Cryptocurrency Influencers automatically and can be considered as a special case of Author Profiling which is the need of the day to expand cryptocurrency market [3].

Crypto Influencers have built a sizable followers on social media sites by expressing their thoughts and observations on cryptocurrencies. However, profiling Cryptocurrency Influencers in social media is very challenging due the very informal kind of data that is available on social media platforms. Given the social media data about the impact the cryptocurrency Influencers have created and the category to which these Influencers belong, profiling the Influencers can be modeled as a text classification problem. However, in a real environment, data collection is a major challenge and real-time profiling needs to be done in a few milliseconds, which implies to process as little data as possible. This demands for the tools which can make predictions accepting very less data.

Conventionally, large amount of annotated data is required to train the ML models. However, collecting, annotating, and validating large data is very expensive. Further, there are many cases where it is just next to impossible to collect large datasets or the available large datasets may not be accessible for public. For example, rare diseases would not have a large number of radiological images or it would be frustrating if smartphones need to have thousands of pictures of users to recognize them and get unlocked. In such scenarios, ML models trained on a small number of samples results in low performance on the Test set. A solution to such scenarios is FSL which aims to build accurate ML models with less training data [4] [5]. FSL which was originally developed for Computer Vision models to work with relatively few training samples is now being used for Natural Language Processing (NLP) applications and the results are encouraging. However, FSL for NLP applications is a young area that needs more research and refinement.

"Profiling Cryptocurrency Influencers with Few-shot Learning" is a shared task[3] in PAN @CLEF 2023 which aims to profile Cryptocurrency Influencers in social media, from a low-resource perspective [6]. Focusing on English Twitter posts, this shared task includes three different subtasks and the details of the subtasks are shown in Table 1. To address the challenges of text classification from a low-resource perspective, in this paper, we describe the FSL models submitted to the above mentioned shared task to profile Cryptocurrency Influencers with very

[2] https://dailycoin.com/crypto-influencers-you-need-to-know-about/

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

Table 1
Details of the subtasks

Subtask	Name of the Subtask	Classes	Description of Label Distribution	Train Set	Test Set
Subtask1	Low-resource Influencer Profiling	Null / No influencer	32 users per label with a maximum of 10 English tweets each	160	220
		Nano			
		Micro			
		Macro			
		Mega			
Subtask2	Low-resource Influencer Interest Identification	Price update	64 users per label with 1 English tweets	320	402
		Technical information			
		Trading matters			
		Gaming			
		Other			
Subtask3	Low-resource Influencer Intent Identification	Subjective opinion	64 users per label with 1 English tweets each	256	292
		Financial information			
		Advertising			
		Announcement			

less data provided by the organizers of the shared task. The proposed models makes use of Google's pre-trained Word2Vec[4] and stsb-bert-base[5] - a Sentence Transformer (ST), to represent text which in turn are used to train LinearSVC models to predict the probabilities of the samples belonging to the classes in the predefined set of classes of a subtask and assign the label of the class having highest probability to the Test sample.

The rest of the paper is organized as follows: Section 2 gives a brief description of the related work followed by the methodology in Section 3. The experiments and results are discussed in Section 4 and the paper concludes with future work in Section 5.

2. Related Work

Profiling Cryptocurrency Influencers and FSL are relatively the new fields for NLP researchers [5]. However, researchers have explored many techniques for Author Profiling. Hence, considering profiling Cryptocurrency Influencers as a special case of Author Pofiling, a brief description of few of the relevant works are given below:

Wang et al. [4] conducted a comprehensive and systematic review of FSL starting from the formal definition of FSL, the relatedness and differences of FSL with relevant learning problems such as weakly supervised learning, imbalanced learning, Transfer Learning (TL) and meta-learning. They focused on learning experiences for small samples and categorised FSL from the perspective of data, model and algorithm. To discuss the pros and cons of each category, they

[4] https://huggingface.co/fse/word2vec-google-news-300
[5] https://huggingface.co/sentence-transformers/stsb-bert-base

explored various learning models including multitask learning, embedding learning, learning with external memory, and generative modelling by refining existing parameters, meta-learned parameters, and optimizer on image data with data augmentation, such as flipping, scaling, rotation, and reflection. Patel et al. [7] presented Sequential Autoregressive Prompting (SAP) - a technique that enables the prompting of bidirectional models. Considering the machine translation task as a case study, they prompt the bidirectional mT5[6] model and demonstrated its few-shot and zero-shot translations. This model outperformed the few-shot translations of unidirectional models like GPT-3[7] and XGLM[8]. Further, they also showed that SAP is a better choice in question answering and text summarization tasks. Their results demonstrate prompt based few-shot and zero-shot learning are emergent techniques in building broader class of language models compared to unidirectional models. Joo and Hwang [8] presented the model submitted to PAN@CLEF 2019 shared task on Author Profiling to determine whether a tweet's author is a bot or human and in case of human, to perform gender identification. Their Gradient Boosted Decision Tree (GBDT) classifier trained by stacking: character count, psycholinguistic features, Term Frequency - Inverse Document Frequency (TF-IDF), Doc2vec and BERT embeddings, obtained accuracies of 0.9333 and 0.8352 for bot identification and gender identification tasks respectively.

Parnami and Lee [5] conducted a detailed survey of the recently proposed FSL algorithms. The agenda of the survey includes learning dynamics, characteristics of FSL, approaches to deal with FSL problems from the perspectives of meta-learning, TL, and hybrid approaches. Further, with open problems and challenges in FSL followed by the discussion about FSL issues, such as training the same way as testing, learning constrained to a single distribution of tasks, performing joint classification from seen and unseen classes, and FSL for domains other than images, they conclude their survey. Tunstall et al. [9] proposed SETFIT (Sentence Transformer Finetuning) - an efficient and prompt-free framework for few-shot fine-tuning of ST. This model first finetunes a pretrained ST on a small number of text pairs to generate rich text embeddings, which are used to train a classification head. This simple framework without prompts or verbalizers have achieved high accuracy with orders of magnitude less parameters than existing techniques. They have also demonstrated that SETFIT can be applied in multilingual settings by simply switching the ST body and evaluated SETFIT with three variants of transformer based models ($\text{SETFIT}_{\text{ROBERTA}}$, $\text{SETFIT}_{\text{MPNET}}$, and $\text{SETFIT}_{\text{MINILM}}$) on various text classification tasks (sentiment analysis, spam detection, and topic classification) considering the available bench-marked datasets (SST-5, AmazonCF, Emotion, EnronSpam, and AGNews). Their proposed $\text{SETFIT}_{\text{ROBERTA}}$ model out performed GPT-3 by exhibiting a maximum average score of 71.3.

Many researchers have worked on Author Profiling considering a substantial amount of data. But, very few works are explored for Author Profiling with few labeled samples. Hence, exploring FSL for profiling Cryptocurrency Influencers with few labeled samples has enough scope and opens up new avenues of research in this topic.

[6] https://huggingface.co/docs/transformers/model_doc/mt5
[7] https://github.com/openai/gpt-3
[8] https://huggingface.co/docs/transformers/model_doc/xglm

3. Methodology

Two models, FSL-Word2Vec and FSL-ST are proposed to address the challenges of the shared task and these models differ primarily in pre-processing and the way of representing texts. The dataset shared by the organizers of the shared task is collected from Twitter [6, 10] and Twitter data will usually be noisy/dirty with lot of URL's, hashtags, user mentions, emojis and numeric information, which are not significant for classification. Hence, text data needs to be cleaned and prepared for text representation. The first step in this direction is to tokenize the input into sentences and applying the pre-processing steps on the sentences depending on the model used for text representation.

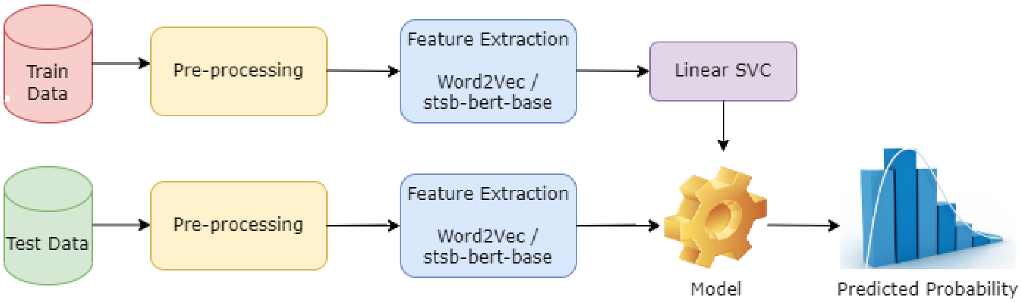

Figure 1: Framework of the proposed methodology

The framework of the proposed methodology is visualized in Figure 1 and the proposed models are described below:

3.1. FSL-Word2Vec

- **Pre-processing** - emojis are converted to the corresponding text and the entire text is lower cased. Further, the noisy content, punctuation, and the stop words are removed from the text. English stopword list[9] available at NLTK library is used to remove the stop words and Porter Stemmer[10] is used to strip the affixes from the words. The remaining content in the pre-processed sentences are given as input to the text representation module.
- **Text Representation** - deals with how efficiently text documents are represented. The introduction of Word2Vec by Mikolov et al. [11] gave rise to the representation of words by a fixed dimension dense vector of small size like 50, 100, 200 and 300. These vectors which are called as pre-trained vectors/embeddings will be trained on a very large corpus. With this representation, any text can be represented as an aggregation of the representation of words. Leveraging pre-trained embeddings for word representation will help to capture the semantic knowledge, even when the task-specific dataset is limited. Google's pre-trained Word2Vec[11] is one such implementation trained on roughly 100 billion words

[9] https://www.nltk.org/search.html?q=stopwords
[10] https://www.nltk.org/api/nltk.stem.porter.html
[11] https://huggingface.co/fse/word2vec-google-news-300

from a Google News dataset. Word2Vec captures the general semantic relationships, patterns, and domains and contains vectors of size 300 for 3 million words and phrases. Using Google's pre-trained Word2Vec model, the text is represented as follows:

  - for each word in a sentence, a vector of 300 dimension is extracted (if the word is not in the vocabulary, a zero vector of length 300 is used)
  - for each sentence, the mean of the vectors of all the words in that sentence is obtained
  - for each text, the mean of the vectors of all the sentences in that text is obtained

  With this arrangement, each text/sample in the dataset will be represented by a dense vector of size 300.

3.2. FSL-ST

- **Pre-processing** - as ST represent the sentences, very limited pre-processing is carried out on the sentences to maintain the sentence structure. This includes converting emojis to the corresponding text and removing the noisy content from each sentence in the given text. Stopword removal, Stemming and converting text to lowercase, are not performed to retain the sentence structure. The pre-processed sentences are given as input to the text representation module.
- **Text Representation** - ST is a Python framework for generating contextualized embeddings for sentences. By embedding sentences into a vector space, ST enables the proximity of similar sentences, allowing for various applications such as semantic search, clustering, and retrieval. With its user-friendly methods, ST simplifies the process of generating embeddings, exhibits state-of-the-arts performance, provides multilingual support, and is reliable as it belongs to open source community. Few-shot and zero-shot approaches have received a great deal of interest in the research community due to the availability of ST and the untapped capacity to use them in resource-constrained domains [9]. stsb-bert-base is a ST model that maps sentences and paragraphs to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search.
  Using stsb-bert-base[12], the text is represented as follows:
  - Each sentence in a text is represented by a vector of 768 dimension
  - for each text, the mean of the vectors of all the sentences in that text is obtained

With this arrangement, each sample in the dataset will be represented by a dense vector of size 768. Hyperparameters and their values of stsb-bert-base are shown in Table 2.

3.3. Model Building

The motivation for using FSL is to learn with few available samples. As the number of training samples for each subtask is very less, the vector representations obtained for the text samples of each subtask are grouped according to the classes/categories and the average of these representations is obtained. This arrangement reduces the number of training samples in the

[12] https://huggingface.co/sentence-transformers/stsb-bert-base

Table 2
Hyperparameters and their values of stsb-bert-base

Hyperparameters	Values
max_seq_length	128
word_embedding_dimension	768
num_hidden_layers	12
hidden_dropout_prob	0.1
attention_probs_dropout_prob	0.1
max_position_embeddings	512

Table 3
Performances of the proposed models

Subtasks	Macro F1 score	
	FSL-Word2Vec	FSL-ST
Subtask1 - Low-resource Influencer Profiling	37.05	**37.92**
Subtask2 - Low-resource Influencer Interest Identification	**46.66**	44.70
Subtask3 - Low-resource Influencer Intent Identification	**50.42**	50.18

subtask to the number of classes in that subtask. These new training samples which are the representative of each class are used to train the LinearSVC model.

To evaluate the model on the Test set, the samples in the Test set will be pre-processed and represented as discussed in Sections 3.1 and 3.2. These vector representations are fed to the LinearSVC model to predict the probabilities of the samples belonging to the classes in the predefined set of classes in the subtask and assign the label of the class having highest probability to the Test samples.

3.4. Experiments and Results

The statistics of the dataset for the subtasks shown in Table 1 indicates that the given datasets are very small in size. If ML classifiers are used in a conventional way, this small size dataset will pose challenges for achieving good performance and generalization on unseen data. This justifies the need for using FSL for profiling Cryptocurrency Influencers.

As the Development set is not provided by the organizers, several experiments were conducted initially by splitting the given Train set into Train and Test set and considering different word representations, ST and classifiers. The combination of word representation - Google's pre-trained Word2Vec, ST - stsb-bert-base and classifier - LinearSVC, which exhibited the best performance in the initial experiments are considered for the actual experiments to obtain the probabilites and predictions on the Test set which in turn are submitted to the organizers for evaluation. The predictions are evaluated based on the macro F1 score as it considers the average F1 score across all classes, providing a comprehensive measure of model effectiveness. Table 3 gives the performances of the proposed models for all the three subtasks. Among the proposed models, FSL-Word2Vec model obtained better macro F1 scores of 46.66 and 50.42 for Subtask2 and Subtask3 respectively whereas FSL-ST model obtained macro F1 score of 37.92 for

Subtask1. The very low macro F1 score for Subtask1 may be because the number of training samples is only 160 which is very small to train any ML model. The performances of both the models for Subtask3 are better compared to that of the other two subtasks. Further, there are no much differences in the performances of the two models in all the subtasks.

4. Conclusion

This paper describes the FSL models proposed for "Profiling Cryptocurrency Influencers with Few-shot Learning" shared task in PAN@CLEF 2023. The proposed methodology consists of two models: i) FSL-Word2Vec - LinearSVC model trained with embeddings extracted from Google's pre-trained Word2Vec and ii) FSL-ST - LinearSVC model trained with sentence embeddings obtained using stsb-bert-base. The performances of both the models for Subtask3 are better compared to that of the other two subtasks. Further, there are no much differences in the performances of the two models in all the subtasks. Efficient FSL techniques will be explored further to handle labeled data with few samples.

References

[1] 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: North American Chapter of the Association for Computational Linguistics, 2022.

[2] U. W. Chohan, Counter-hegemonic finance: The gamestop short squeeze, in: Financial Crises eJournal, 2021.

[3] M. Chinea-Rios, T. Müller, G. L. D. la Pena Sarrac'en, F. Rangel, M. Franco-Salvador, Zero and few-shot learning for author profiling, in: International Conference on Applications of Natural Language to Data Bases, 2022.

[4] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, in: ACM Comput. Surv., volume 53, 2021, pp. 63:1–63:34. URL: https://doi.org/10.1145/3386252. doi:10.1145/3386252.

[5] A. Parnami, M. Lee, Learning from few examples: A summary of approaches to few-shot learning, in: ArXiv, volume abs/2203.04291, 2022.

[6] 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.

[7] A. Patel, B. Li, M. S. Rasooli, N. Constant, C. Raffel, C. Callison-Burch, Bidirectional language models are also few-shot learners, in: ArXiv, volume abs/2209.14500, 2022.

[8] Y. Joo, I. Hwang, Profiling on social media : An ensemble learning model using various features notebook for pan at clef 2019, 2019.

[9] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient Few-shot Learning without Prompts, in: arXiv preprint arXiv:2209.11055, 2022.

[10] 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.

[11] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013.