Department: Affective Computing and Sentiment Analysis
Editor: Erik Cambria, Nanyang Technological University

Sentiment Classification of Cryptocurrency-Related Social Media Posts

Mikolaj Kulakowski
Erasmus University Rotterdam

Flavius Frasincar
Erasmus University Rotterdam

***Abstract*—Many researchers agree that sentiment analysis can improve the performance of quantitative trading models. We develop two off-the-shelf solutions for analysing the sentiments of cryptocurrency-related social media posts. First, we post-train and fine-tune a Twitter-oriented model based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, BERTweet, on the cryptocurrency domain resulting in CryptoBERT. Second, we generate the Language-Universal Cryptocurrency Emoji (LUKE) sentiment lexicon and prediction pipeline, utilising the sentiment of emojis prevalent in social media. CryptoBERT is highly accurate, while LUKE is suitable for non-English posts, thus allowing for direct classification and noisy label generation in less popular languages. Our research can help cryptocurrency investors develop trading software supported by sentiments mined from social media.**

CRYPTOCURRENCY trading is a growing field in finance, with a market capitalization exceeding $1 trillion[1]. Cryptos are also increasingly present in social media, motivating investors to utilise online sentiments to improve their cryptocurrency trading algorithms. Such aggregation can be achieved through Sentiment Analysis (SA), a growing field in Natural Language Processing (NLP) that extracts the affective meaning and sentiment polarity from text. Within the financial context, an SA model takes a text as input and returns a Sentiment Score (SS) that can be either bullish (positive), neutral, or bearish (negative). The resulting SS can be applied in a trading model, improving financial forecasting [1].

In our approach, we develop two off-the-shelf solutions for classifying the sentiments of cryptocurrency-related social media posts. First, we post-train and fine-tune a model based on the architecture of the Bidirectional Encoder Representation from Transformers (BERT) model [2]. BERT is a state-of-the-art language model, able to utilise a huge corpus of data to learn the

[1]http://coinmarketcap.com/

This article is part of the IEEE IS ACSA series

numerical representations of texts from a given language domain. Starting from a Twitter-based model, BERTweet [3], we train it further on a corpus of cryptocurrency-related social media posts. We name the resulting model CryptoBERT.

Furthermore, we pursue a cross-lingual model, able to train on English-language posts and predict sentiments for other languages. We thus turn our attention towards emojis, which appear on social media platforms in widely unchanged forms among users' languages. Emojis are found to keep their SS regardless of the language used in communication [4], while sentences in different languages, with similar emojis, carry similar emotional information [5]. Thus, much emotion is preserved in emojis when text language is changed. Last, emojis and emoticons have a dominating effect on sentiment polarity on paragraph level [6]. For our second approach, we thus automatically generate an emoji sentiment lexicon that could work as a bridge to text written in less-used languages. We call our solution the Language-Universal Cryptocurrency Emoji (LUKE) lexicon. Furthermore, a prediction pipeline is developed to extract the SS of texts using the LUKE lexicon. LUKE entries comprise single emojis, as well as emoji pairs.

The rest of the paper is structured as follows. First, the data sets used are presented. Next, we describe the methods used in generating each off-the-shelf solution. Further, the performance of each solution is evaluated. The paper ends with the conclusion of our work. The code used for creating and using the CryptoBERT and LUKE solutions, as well as the LUKE lexicon itself, can be found on https://github.com/mikik1234/CryptoBERT-LUKE, while the best performing CryptoBERT model can be downloaded from https://huggingface.co/ElKulako/cryptobert.

DATA

The cryptocurrency social media corpus used in our research consists of 3.207 million posts, including 496 thousand from Twitter (twitter.com) collected from 2018-07-11 to 2018-07-24, 172 thousand from Reddit (reddit.com) collected between 2021-05-01 and 2022-04-30, 664 thousand from Telegram (telegram.org) collected from 2020-11-16 to 2021-01-30, and 1.875 million from StockTwits (stocktwits.com) collected from 2021-11-01 to 2022-06-30. The StockTwits posts are labelled by their authors as either bullish or bearish; we assume neutral sentiment if no label is assigned. The corpora from other sources are unlabelled, therefore only StockTwits posts are used for supervised training and evaluation, while the other sources are only used for the unsupervised post-training of the CryptoBERT model. Last, due to differences in language used for various cryptocurrencies, for classification, we only consider StockTwits posts about the three most discussed currencies, Bitcoin (BTC.X), Ethereum (ETH.X), and Shiba Inu (SHIB.X) while performing supervised training. The StockTwits training set ranges from 2021-11-01 to 2022-06-15 and contains 1.332 million posts. For model evaluation, the StockTwits test set consists of 83,257 posts collected from 2022-06-16 to 2022-06-30.

Preceding training, all corpora undergo a cleaning procedure. For each post, we remove the Chinese, Japanese, and Korean letters, crypto wallet addresses, URLs, cashtags ($), hashtags (#), usernames (@), and retweets (RT). Then, we fix known special character encoding errors, multiple dots and spaces. Last, we convert all characters to lowercase, and remove duplicate posts and posts containing less than four words.

Additional data filtering must be performed for the LUKE lexicon. First, we extract posts with at least one emoji from StockTwits data. Additionally, the emoji training set is further filtered, by only considering posts with emojis that often appear in either bullish or bearish setting. We also exclude emojis that are often found in all three sentiment classes. We consider 91,758 posts, from 2021-11-01 to 2022-04-30, for the emoji training set used in constructing the LUKE sentiment lexicon; while using an emoji validation set of 20,761 posts from 2022-05-01 to 2022-06-15 to fine-tune the LUKE prediction pipeline. The emoji test set uses 11,984 posts from 2022-06-16 to 2022-06-30.

The StockTwits data set is at https://huggingface.co/datasets/ElKulako/stocktwits-crypto, while CryptoBERT post-training corpus is under https://huggingface.co/datasets/ElKulako/cryptobert-posttrain. The StockTwits emoji data set is at https://huggingface.co/datasets/ElKulako/stocktwits-emoji.

METHODOLOGY

We propose two sentiment classification methods for cryptocurrency-related social media posts. The first method is based on post-training and fine-tuning a model built using the BERT architecture [2]. The second method uses SVM models to classify emojis as either bullish or bearish, which are then assigned to the LUKE sentiment lexicon.

CryptoBERT

The procedure used in training the CryptoBERT model is described below. The BERTweet model [3] is used as the starting step to be further post-trained on our social media crypto corpus. For post-training, we perform the masked language modelling (MLM) training task, based on the robustly optimized BERT pretraining approach (RoBERTa) [7]. Both RoBERTa and BERTweet use a byte-level byte-pair encoding (BPE) tokenizer [8] to convert inputs into numerical representations of a vocabulary, called tokens. The MLM task focuses on masking roughly 15% of input tokens, which are used as targets for model predictions based on context. MLM can be performed in an unsupervised setting, using our entire set of 3.207 million posts.

Following the original BERT procedure, we first train the BERTweet model on a shorter sequence length of 32 tokens and then set the sequence length at 128 tokens. Inspired by [7], we also introduce multiple masking in our training. The weights are optimized with Adam. We train for 120 epochs, with 10 different masks (12 epochs per mask), with a max sequence length of 32, and for 12 epochs at a sequence length of 128. The resulting model is CryptoBERT.

To fine-tune CryptoBERT for the sentiment classification task, we use the StockTwits training set. Furthermore, since there is a high imbalance among sentiment classes, we introduce sampling to even out the training sets. Namely, we consider undersampling and oversampling. First, we use the bearish set size of 124,451 posts and sample without replacement from the other two classes, reaching 124,451 posts per class. This training set of 373,353 posts is used to fine-tune and evaluate all classifiers. Additionally, to maximise performance, we perform oversampling of the training data, by sampling with replacement from the (smaller) bearish and neutral sets, so that all three classes have 676,701 posts, corresponding with the size of the largest (bullish) set. This set of 2.03 million posts is used to fine-tune the CryptoBERT model and the best-performing benchmark, BERTweet. These receive an "XL" label, thus resulting in CryptoBERT XL and BERTweet XL. For both sets, 10% of training data is set aside for validation.

LUKE Sentiment Lexicon

The procedure used for constructing the LUKE sentiment lexicon is described below. We train SVM classifiers, with emojis as features, which are then used to classify single emojis and emoji pairs as either bullish or bearish.

First, due to a large class discrepancy, with only 7 thousand bearish against 57 thousand bullish posts in the emoji training set, we train two separate SVM classifiers. The first model is trained to distinguish posts as either bearish or "not bearish" (bullish or neutral). In this step, we use 20 thousand bearish posts sampled with replacement and 20 thousand "not bearish" posts sampled without replacement. The second set is used for classifying posts as either bullish or "not bullish" (bearish or neutral). It contains 40 thousand bullish posts sampled without replacement and 40 thousand bearish and neutral posts sampled with replacement. As such, the bullish and bearish parts of the LUKE lexicon are generated independently. Last, in both sets, 10% of training data is set aside for SVM hyperparameter optimisation. We use a second-degree polynomial kernel in both SVMs.

Second, we create lists of the most prevalent emojis in the bullish (bearish) class, which are combined into pairs (class-wise). Next, the SVMs are used to predict the probability of being in their respective class, for every single emoji and emoji pair. If a given forecast has a probability that exceeds 0.6, that item is added to the LUKE lexicon.

Last, the SS for each entry is calculated by

$$SS_i = (-1)^{I(bearish_i)}(9P(i) - 5), \quad (1)$$

where $P(i)$ is the class probability of emoji i, while the indicator function $I(bearish_i)$ returns 1 if a given entry is bearish and 0 otherwise. These scores are then added to LUKE entries.

Since the probabilities range from 0.6 to 1, the resulting SS range from -4 to 4, following the entries of VADER. The lexicons could thus be combined, to enhance VADER's handling of emojis. We exclude SS values from -0.4 to 0.4, which are eliminated to avoid noisy lexicon entries.

LUKE Prediction Pipeline

We develop the LUKE prediction pipeline to forecast the sentiments of posts. First, we check each post for LUKE emoji pairs and sum their SS; if no pairs are found, we sum the SS of LUKE single emojis found in the post. Our reasoning is that the pairs have a stronger influence, while also changing the meaning of individual emojis. If the aggregate score exceeds our threshold of ± 1.4, the post is classified as either bullish (positive score) or bearish (negative score). Otherwise, we treat it as neutral. The LUKE prediction pipeline is fine-tuned using the emoji validation set.

Last, to evaluate whether pairs are a valuable addition, we also consider a case where only single emojis are used. This method provides much faster forecasts.

Evaluation

The following benchmarks are considered for evaluation. First, VADER [9] lexicon, since it is state-of-the-art in the social media domain. Next, three BERT-like models are considered, specifically the generic BERT [2], the FinBERT model, trained on the financial domain [10], and the BERTweet model [3]. All models are fine-tuned on the undersampled StockTwits training set. Additionally, the best-performing benchmark, BERTweet, is also fine-tuned on the oversampled StockTwits training set, resulting in BERTweet XL.

RESULTS

The performance measures of BERT-based models are displayed in Table 1 for the StockTwits test set. It can be seen that the post-trained CryptoBERT model outperforms the other methods on both the undersampled (CryptoBERT) and oversampled (CryptoBERT XL) training sets. The best-performing model overall is CryptoBERT XL with an accuracy of 58.49% and macro F_1 score of 58.83%. Among our benchmarks, BERTweet provides the best performance, while the VADER lexicon has the least accurate forecasts. Both methods were created using Twitter inputs, implying how an adjustable solution is beneficial in a new domain. Last, for both fine-tuning samples, CryptoBERT consistently outperforms BERTweet, implying that the domain adaptation of BERT-based models using unlabelled training data improves their forecasting ability.

Table 1. Performance of sentiment classifiers on the StockTwits test set.

Model	Accuracy	F_1 score	Precision	Recall
VADER	37.10	36.53	33.81	37.02
BERT	53.66	53.94	45.71	58.23
FinBERT	52.74	52.79	43.50	57.69
BERTweet	55.29	55.45	46.39	59.73
CryptoBERT	**55.60**	**55.79**	**46.58**	**60.44**
CryptoBERT XL	**58.49**	**58.83**	**51.98**	**61.37**
BERTweet XL	58.07	58.39	51.28	61.08

Table 2. Performance of sentiment classifiers on the emoji test set.

	Accuracy	F_1 score	Precision	Recall
LUKE single emojis	48.80	48.02	45.77	51.05
LUKE with pairs	48.77	48.16	46.55	50.93
VADER	35.87	36.10	36.80	35.42
BERT	55.04	50.44	41.96	55.88
FinBERT	54.06	47.17	38.34	54.27
BERTweet	60.03	55.22	46.08	59.80
CryptoBERT	**60.31**	**55.79**	**46.59**	**60.60**
CryptoBERT XL	62.35	**59.49**	**52.94**	61.15
BERTweet XL	**62.52**	59.31	52.19	**61.27**

The performance of LUKE on the emoji test set is displayed in Table 2. First, it can be seen that LUKE outperforms the VADER lexicon, thus using emojis as lexicon entries is beneficial to only considering their textual descriptions, as VADER does. Next, incorporating emoji pairs in the forecast does not provide improvements over using single emojis. Furthermore, incorporating emoji pairs causes a substantial increase in the time required to make predictions.

CONCLUSION

In this paper, we presented two off-the-shelf solutions for analysing the sentiments of cryptocurrency-related social media posts. Our first approach uses a BERT-like model trained on a large corpus of Twitter data, BERTweet, and post-trains and fine-tunes it using a cryptocurrency corpus. The resulting model, CryptoBERT, delivers accurate performance in the sentiment classification of cryptocurrency text. For our second approach, we construct the LUKE sentiment lexicon, which outperforms VADER on the task

of cryptocurrency sentiment classification.

Thus, if one wishes to maximize sentiment prediction accuracy for a cryptocurrency trading model, CryptoBERT performs best. That said, the LUKE sentiment lexicon and prediction pipeline can be employed towards other tasks, such as forecasting distant labels for non-English corpora, so that their information can be used directly in sentiment classification, or as inputs in the training of other models.

REFERENCES

1. F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural Language Based Financial Forecasting: A Survey," *Artificial Intelligence Review*, vol. 50, no. 1, 2018, pp. 49–73.
2. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, ACL, 2019, pp. 4171–4186.
3. D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020)*, ACL, 2020, pp. 9–14.
4. Z. Chen et al., "Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification," *Proceedings of the 30th World Wide Web Conference (WWW 2019)*, ACM, 2019, p. 251–262.
5. N. Choudhary et al., "Contrastive Learning of Emoji-Based Representations for Resource-Poor Languages," *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018)*, Lecture Notes in Computer Science, Springer, 2018, pp. 129–141.
6. A. Hogenboom et al., "Exploiting Emoticons in Polarity Classification of Text," *Journal of Web Engineering*, vol. 14, 2015, pp. 22–40.
7. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
8. R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, ACL, 2015, pp. 1715–1725.
9. C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*, The AAAI Press, 2014, pp. 216–225.
10. D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv preprint arXiv:1908.10063*, 2019.

Mikolaj Kulakowski is a graduated MSc in Quantitative Finance at Erasmus School of Economics. His interests include NLP, cryptocurrencies, and trading systems. Contact him at kulakm@student.eur.nl.

Flavius Frasincar is an assistant professor at Erasmus University Rotterdam. His research interests lie in Web information systems, personalization, machine learning, sentiment mining, and the Semantic Web. He was awarded a PhD in information systems from Eindhoven University of Technology. Contact him at frasincar@ese.eur.nl.