Sentiment Classification of Cryptocurrency-Related Social Media Posts

Mikolaj Kulakowski
Erasmus University Rotterdam

Flavius Frasincar
Erasmus University Rotterdam

***Abstract*—There is a rising demand for optimal cryptocurrency trading algorithms. Many researchers agree that sentiment analysis can improve the performance of quantitative trading models. This paper focuses on creating off-the-shelf solutions for analysing the sentiments of cryptocurrency-related social media posts. We develop two solutions. 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 a prediction pipeline using the Support Vector Machine (SVM), which classifies posts based on aggregating the sentiment of the emojis that they contain. Though less accurate than CryptoBERT, LUKE is suitable for non-English posts, thus allowing for immediate 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 total market capitalization exceeding $1 trillion[1]. Aside from being widely traded, cryptos are also increasingly present in social media. Indeed, on average, the bitcoin hashtag is mentioned in over one hundred thousand tweets per day[2]. Knowing the above, an investor may wish to utilise the abundant social media information to improve their cryptocurrency trading model. Such aggregation can be achieved through Sentiment Analysis (SA), a growing field in Natural Language Processing (NLP) tasked with extracting the affective meaning and sentiment polarity from text [3]. In 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 then be applied in a model, which has been shown to improve financial forecasting in many cases [19]. Thus, a need arises for an off-the-shelf SA solution that can aggregate the emotions conveyed in cryptocurrency posts.

[1]http://coinmarketcap.com/
[2]http://bitinfocharts.com/comparison/bitcoin-tweets.html

In our approach, we focus on developing 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 [7]. It is a state-of-the-art language model, able to utilise a very large corpus of data to learn the numerical representations of texts from a given language domain. We name the resulting model CryptoBERT. Furthermore, we are also interested in 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 come from Japanese words "e" (picture) and "moji" (character) and are pictograms widely used in electronic communication to provide emotional cues within text.

They appear on social media platforms in widely unchanged forms among users' languages; more importantly, the language-universal property is largely found in emojis [5], where they keep their SS regardless of the language used in communication. 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. A sentiment lexicon is a resource that contains influential terms along with their SS [8].

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. Last, we investigate whether including pairs of emojis in the lexicon can improve prediction accuracy, over using only a "bag" of single emojis. Our reasoning is that sometimes two emojis appearing together can influence each other's meaning, for instance, the "bear" emoji is traditionally bearish, but combined with the "sweat droplets" emoji could be translated into "bears are sweating", which means that the pessimists are unhappy, thus gaining positive meaning as a result of the combination of two negatives. The rest of the paper is structured as follows. First, we discuss our research's related work and scientific relevance. Then, the data sets used are presented for each performed task. 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 and future research directions. 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.

RELATED WORK AND SCIENTIFIC RELEVANCE

Much research has been conducted on applying SA in the financial domain. For instance, in [13], a stock prediction is performed by incorporating the numerical features and market-driven news sentiments of target stocks and the sentiment of their related stocks. Further, various machine learning techniques are used by [14] to determine the best way of allocating stock portfolios. The authors demonstrate that incorporating financial sentiment improves the performance of optimal wealth allocation. Regarding cryptocurrency trading, [17] shows that public sentiment can be used to predict Bitcoin trading volume with state-of-the-art accuracy, especially when combining textual and non-textual features, such as likes and retweets. Last, the investigation of [16] compares the benefits of SS as additional features to be used in constructing stock and cryptocurrency trading algorithms. The authors find that the improvements are more pronounced for cryptos than for stocks. One can notice the benefit of SA in financial models, however, most solutions focus on finance-specific language instead of the cryptocurrency domain; thus, a need for a bespoke solution emerges.

Investigating possible solutions, one may turn their attention to the language model BERT [7], developed by Google. It is based on transformer architecture and can be trained on a large, unsupervised corpus of text, to learn the characteristics of a given language domain. The ease of domain adaptation has resulted in many models being trained for a particular domain; for example, the FinBERT model [2] has been trained on a large corpus of financial news and reports to analyse the sentiments of financial texts. It has been widely applied in financial models, even used by [17] to predict the bitcoin trading volume; we believe that this task could be further enhanced by a cryptocurrency-specific method. Furthermore, the social media domain often uses the BERTweet model [15], which was trained on a large corpus of Twitter posts. The BERT-based models have demonstrated state-of-the-art performance in the sentiment classification task. Last, the Support Vector Machine (SVM) models are widely used in text mining approaches and SA [1], as they handle sparse data very well [4]. Regarding social media lexicons, the current state-of-the-art is the Valence-Aware Dictionary and sEntiment Reasoner (VADER) lexicon, manually created by a group of experts, based on terms used in a corpus of social media posts [10]. It is both a sentiment lexicon and a prediction pipeline, with the SS values ranging from -4 to 4. However, VADER's handling of emojis is limited to converting them into text using their Unicode descriptions. It has been shown that pictograms such as emojis or emoticons often dominate the text's sentiment polarity on the paragraph level [9].

Last, emojis can often capture a meaning shared across languages, especially in case of related cultures. For instance, [6] asserts that sentences in different languages, with similar emojis, carry similar emotional information. Thus, much emotion is preserved in emojis even when the target language is changed. Moreover, emojis can be used to predict cross-lingual sentiment, using a model trained on a resource-rich language to forecast sentiments for a resource-poor language [5], [20]. Therefore, the construction of an emoji-specific sentiment lexicon appears beneficial.

DATA

The cryptocurrency social media text corpus used in our research consists of 3.207 million posts, including 496 thousand posts 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 posts 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 BERT-based 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). The StockTwits training set ranges from 2021-11-01 to 2022-06-15. For model evaluation, the StockTwits test set consists of posts collected from 2022-06-16 to 2022-06-30.

Before the data is passed to our methods, all corpora undergo a cleaning procedure. First, the duplicate and empty posts are removed. Next, 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 (e.g., for the ampersand or apostrophe), multiple dots are replaced with triple dots, while multiple spaces are replaced with single spaces. All letters are also converted to lowercase, as the BERT-based models are case-sensitive. Last, we again remove the duplicate posts and delete all posts containing less than four words.

Further, additional data filtering must be performed for the LUKE emoji sentiment lexicon and prediction pipeline. First, we extract a subset of the StockTwits data set, containing only the posts that have at least one emoji. Additionally, for the training data, one may wish to minimize the number of emojis to build the lexicon from. Thus, 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 very prevalent in all three sentiment classes. On the other hand, to tune and test the coverage and performance of the LUKE lexicon's prediction pipeline, we use all posts that contain emojis. Moreover, aside from the test set, a validation set is needed to tune and optimize the LUKE prediction pipeline. Therefore, we only consider data from 2021-11-01 to 2022-04-30 for the emoji training set used in constructing the LUKE sentiment lexicon, while using the posts from 2022-05-01 to 2022-06-15 as a validation set for the LUKE sentiment prediction pipeline. The emoji test set uses StockTwits posts from 2022-06-16 to 2022-06-30.

The complete StockTwits training data set contains 676,701 bullish, 530,545 neutral, and 124,451 bearish posts, while the test set contains 24,572 bullish, 37,758 neutral, and 20,927 bearish posts. This amounts to 1.415 million observations, including 1.332 million posts in the StockTwits training data and 83,257 posts in the StockTwits test set. Regarding the emoji data set, the emoji training data contains 91,758 observations, including 57,932 bullish, 26,516 neutral, and 7310 bearish posts. The emoji validation set has 20,761 observations, with 9143 bullish, 7534 neutral, and 4084 bearish posts. The emoji test set consists of 5172 bullish, 4199 neutral, and 2613 bearish posts, amounting to 11,984 examples. The emoji data set has thus 124,503 posts in total, including 72,247 bullish, 38,249 neutral, and 14,007 bearish posts.

METHODOLOGY

This paper proposes 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 [7]. 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 displayed in Figure 1. First, the BERTweet model [15] is used as the starting step to be further post-trained on our social media crypto corpus. For post-training, we follow a combination of steps used in the original BERT algorithm, and of the steps used in the robustly optimized BERT pretraining approach (RoBERTa) [12]. Both RoBERTa and BERTweet use a byte-level byte-pair encoding (BPE) tokenizer [18] to convert inputs into numerical representations of a 64-thousand-term vocabulary, called tokens. Further, in the RoBERTa framework, only the masked language modelling (MLM) training task is used. The task focuses on masking roughly 15% of input tokens, which are then used as targets for model predictions based on their context. Such a procedure can be performed in an unsupervised setting, thus allowing the use of our entire data 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. Since less than 20% of posts exceed 32 tokens, we can preserve most of the training accuracy, while severely limiting the time required for training. Inspired by the methods presented in [12], we also introduce multiple masking in our training. Namely, for the training part with a maximum sequence length of 32, we make 10 copies of data, thus allowing for 10 different ways each post can be masked. Only one mask is applied for the sequence length of 128. The weights are optimized with Adam [11]; we train for 120 epochs (12 epochs per mask) using the length of 32 and for 12 epochs at a sequence length of 128. The resulting model is CryptoBERT, which takes a piece of text as input and returns a sequence of 768-dimensional vectors representing the sentence and each individual token.

To fine-tune CryptoBERT for the sentiment classification task, we use the StockTwits training data set of 1.332 million labelled posts about Bitcoin, Ethereum, and Shiba Inu coins. Furthermore, since there is a high imbalance among sentiment classes, we introduce sampling to make our training sets more even. Namely, we consider undersampling and oversampling, since the more sophisticated methods are not directly compatible with textual data. First, we use the bearish set size of 124,451 posts and sample without replacement from the other two classes, so there are 124,451 posts per class. This training set of 373,353 posts is used to fine-tune every BERT-based model, with 10% of data being set aside for validation. Additionally, to better use our information set, 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) class training set. This set of 2.03 million posts is then used to fine-tune the CryptoBERT model and the best-performing benchmark, the BERTweet model. These two models are then given an "XL" label, thus resulting with CryptoBERT XL and BERTweet XL. Again, 10% of training data is set aside for validation. The models trained on the undersampled set are used for comparison. In contrast, the oversampled training set is used to maximize model performance, so cryptocurrency investors can use the best model available.

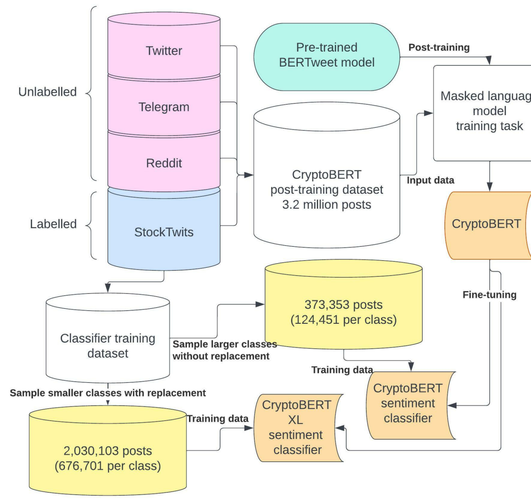

Figure 1. CryptoBERT post-training and fine-tuning flowchart.

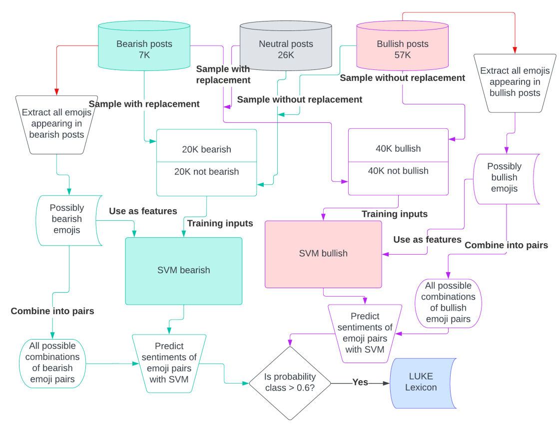

Figure 2. LUKE lexicon generation procedure.

LUKE Sentiment Lexicon

The procedure used for constructing the LUKE sentiment lexicon is shown in Figure 2. The crux of our approach is training 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, with emojis as features, to better use the information set. 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" (bullish or neutral) 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 model validation. We use a second-degree polynomial kernel in both SVMs, as determined through validation. Second, we create lists of possibly bullish (bearish) emojis, based on the most prevalent emojis in the bullish (bearish) class, provided that they are sufficiently rare

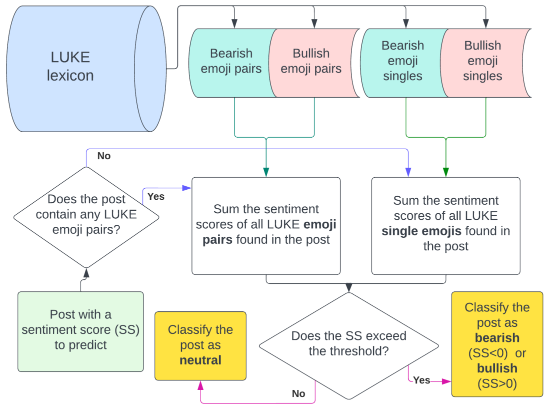

Figure 3. LUKE lexicon prediction pipeline.

among the posts of the opposing class. These possibly bullish (bearish) emojis are combined into pairs (within a given sentiment class). Next, the two 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 (determined based on the emoji validation set), the corresponding pair (single) is added to the LUKE lexicon.

Last, to make LUKE sentiment scores compatible with those of VADER, the SS for each entry is calculated by

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

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, so that they can be used while making predictions. Since the probabilities range from 0.6 to 1, the resulting SS range from -4 to 4, following the entries of VADER. One may thus use the lexicons together, to enhance VADER's handling of emojis. Last, one may note that the above approach excludes SS values from -0.4 to 0.4. Since we are using a machine learning approach, which is not as precise as a group of experts, it is difficult to distinguish whether the numbers spanned by this interval are indeed bearish or bullish. Thus, they are excluded to avoid noisy lexicon entries.

LUKE Prediction Pipeline

Having constructed the LUKE lexicon, the next step is using the found emojis to forecast the sentiments of posts from the emoji test set. The LUKE prediction pipeline is shown in Figure 3. First, we check whether a given post contains any LUKE emoji pairs; if it does, then only the SS of the pairs are counted for the post. Our reasoning is that the pairs have a stronger influence together, while also changing the meaning of individual emojis. If no pairs are found, we consider all LUKE single emojis found in the post. In either case, the SS of LUKE entries are added together, to calculate the SS of the post. If the score exceeds our threshold of ± 1.4 (determined based on the emoji validation set), the post is classified as either bullish (positive score) or bearish (negative score). Otherwise, we treat it as neutral.

Last, to evaluate whether pairs are a valuable addition, we also consider a case, where only single emojis are used. In this approach, the search for emoji pairs is forsaken, while only the SS of single emojis are considered for each post. This method provides much faster forecasts, thus, it is beneficial for cases where time is essential for a crypto investor. The LUKE prediction pipeline is fine-tuned using the emoji validation set.

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

This table presents the performance measures for classification on the StockTwits test set. All values are in percentages. The last two rows correspond to the classifiers trained on the larger, sampled data set. The best scores are in bold for each measure. The results of the last two rows are analysed separately; their measures are additionally put in bold if they outperform the classifiers trained on the smaller training set.

Evaluation

In order to establish the performance of our solutions, a number of benchmark models are considered. First, the VADER lexicon and prediction pipeline [10] are used, since this lexicon is a state-of-the-art sentiment lexicon in the social media domain. Next, three BERT-like models are considered, specifically the generic BERT-base-uncased model [7], the FinBERT model, trained on the financial domain [2], as well as the BERTweet model trained on the Twitter corpus [15]. All the BERT-based models are fine-tuned on the same undersampled StockTwits training data set as CryptoBERT. Additionally, the best-performing benchmark method, BERTweet, is also fine-tuned on the large, oversampled StockTwits training set, resulting with the model named "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 StockTwits data sets. The best-performing model overall is CryptoBERT XL with an accuracy of 58.49% and macro F_1 score of 58.83%, closely followed by BERTweet XL; demonstrating that training on the larger set of data is beneficial for forecasting performance. Among our benchmarks, BERTweet provides the best performance with respect to all measures used. The VADER lexicon has the least accurate forecasts, with an accuracy of 37.10% and the macro F_1 score of 36.53%. Both methods were created using Twitter inputs, however, BERTweet could also be fine-tuned on our data, implying how an adjustable, automatic solution is beneficial when a new domain is considered. Furthermore, the FinBERT model performs worse than the other BERT-based models, including the original BERT model, implying that there is a substantial difference in language used between the financial and cryptocurrency investor domains. 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.

The performance of the LUKE sentiment lexicon and prediction pipeline is displayed in Table 2, compared with all the other models presented in Table 1. They are evaluated on the emoji test set. 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, one may see that incorporating emoji pairs in the forecast does not provide noticeable improvements, compared to only using a "bag" of single emojis, to make predictions for a given post. Furthermore, since there are over 23,000 influential emoji pairs, while there are only around three hundred single emojis, incorporating emoji pairs causes a substantial increase in time required to make predictions, thus slowing down the pipeline.

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**

This table presents the performance measures for classification on the emoji test set. All values are in percentages. The first row reports measures for using only single emojis from LUKE. The second row from using both singles and emoji pairs. The last two rows correspond to the classifiers trained on the larger, sampled data set. The best scores are in bold for each measure. The results of the last two rows are analysed separately; their measures are additionally put in bold if they outperform the classifiers trained on the smaller training set.

Next, looking again at Table 2 one may see that all the BERT-based models outperform the LUKE lexicon. However, the BERT-based models can only be used for English-language inputs, while LUKE can make predictions for any data set that uses the Unicode emojis. Last, when comparing performance measures between Table 1 and Table 2 one may see that the BERT-based models are more accurate when the emoji-rich data is considered, while the accuracy of VADER deteriorates when only the posts with emojis are considered. This implies that emojis provide beneficial information that BERT-based models are able to incorporate, while VADER's handling of emojis is unsuitable for cryptocurrency data. We thus infer that the words used in emoji descriptions rarely have respective connotations in the world of cryptocurrency investing.

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 both labelled and unlabelled cryptocurrency corpora. The resulting model, CryptoBERT, delivers state-of-the-art performance in the sentiment classification of cryptocurrency text. For our second approach, we train two SVM classifiers to construct the LUKE sentiment lexicon, which outperforms VADER on the task of cryptocurrency sentiment classification of StockTwits posts. 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 indirectly, as inputs in the training of other models. Last, we find that using emoji pairs in combination with single emojis does not lead to performance improvements, while costing a substantial amount of time compared to using just the single emoji entries.

For further research, if reaching higher accuracy in the models is desirable, we suggest pre-training a CryptoBERT model from scratch using a larger corpus of cryptocurrency-related text with its own bespoke vocabulary. Another research suggestion would be to combine the power of LUKE and VADER, using the latter's understanding of written text and the former's superior handling of emojis.

REFERENCES

[1] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed. Sentiment analysis using SVM: A systematic literature review. International Journal of Advanced Computer Science and Applications, 9(2), 2018.
[2] D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019.
[3] E. Cambria. Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2):102-107, 2016.
[4] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In Proceedings of ICLM, pages 96-103. ACM, 2008.
[5] Z. Chen, S. Shen, Z. Hu, X. Lu, Q. Mei, and X. Liu. Emoji-powered representation learning for cross-lingual sentiment classification. In Proceedings of the 30th World Wide Web Conference (WWW 2019), page 251262. ACM, 2019.
[6] N. Choudhary, R. Singh, I .Bindlish, and M. Shrivastava. Contrastive learning of emoji-based representations for resource-poor languages. In Proceedings of CICLing, Lecture Notes in Computer Science, pages 129-141. Springer, 2018.
[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171-4186. ACL, 2019.
[8] B. Heerschop, A. Hogenboom, and F. Frasincar. Sentiment lexicon creation from lexical resources. In Proceedings of the 14th International Conference on Business Information Systems, volume 87 of Lecture Notes in Business Information Processing, pages 185-196. Springer, 2011.
[9] A.Hogenboom,D.Bal,F.Frasincar,M.Bal,F.deJong, and U. Kaymak. Exploiting emoticons in polarity classification of text. Journal of Web Engineering, 14:22-40, 2015.
[10] C. Hutto and E. Gilbert. VADER: A parsimonious rulebased model for sentiment analysis of social media text. In Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014), pages 216225. The AAAI Press, 2014.
[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 2015.
[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
[13] Y. Ma, R. Mao, Q. Lin, P. Wu, and E. Cambria. Multisource aggregated classification for stock price movement prediction. Information Fusion, 91:515-528, 2023.
[14] L. Malandri, F. Z. Xing, C. Orsenigo, C. Vercellis, and E. Cambria. Public mooddriven asset allocation: the importance of financial sentiment in portfolio management. Cognitive Computation, 10(6):1167-1176, 2018.
[15] D. Q. Nguyen, T. Vu, and A. T. Nguyen. BERTweet: A pre-trained language model for English tweets. In Proceedings of EMNLP, pages 9-14. ACL, 2020.
[16] M. Ortu, N. Uras, C. Conversano, G. Destefanis, and S. Bartolucci. On technical trading and social media indicators in cryptocurrencies? price classification through deep learning. Expert Systems with Applications, 198:116804, 2022.
[17] J. Saha, S. Patel, F. Xing, and E. Cambria. Does social media sentiment predict bitcoin trading volume? In Proceedings of ICIS, 2022.
[18] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In Proceedings of ACL, pages 1715-1725. ACL, 2015.
[19] F. Z. Xing, E. Cambria, and R. E. Welsch. Natural language based financial forecasting: A survey. Artificial Intelligence Review, 50(1):49-73, 2018.
[20] J. Zhang, T. Liang, M. Wan, G. Yang, and F. Lv. Curriculum knowledge distillation for emoji-supervised cross-lingual sentiment analysis. In Proceedings of the 27th Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), pages 864-875. ACL, 2022.

Mikolaj Kulakowski is a graduated Master of Science student at Erasmus University Rotterdam. His interests include natural language processing, cryptocurrencies, and automated trading systems. He received an MSc in Quantitative Finance from Erasmus School of Economics. 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.