Using BERTweet to Perform Sentiment Analysis on the Crypto Market

Marcos Geraldo and Thomas Claiborne

NLP - DATASCI 266

University Of California, Berkeley - MIDS

Abstract

The influence of social media sentiment on financial markets has become increasingly evident, particularly in volatile sectors like cryptocurrency. This paper presents a sentiment classification framework designed to detect financially relevant sentiment in crypto-related tweets using transformer-based language models. We use BERTweet-large, a model pretrained on Twitter data, and apply Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. Tweets are preprocessed, filtered for financial relevance, and labeled using a hybrid heuristic and score-based approach to distinguish between positive, neutral, and negative sentiment. Our experiments demonstrate that LoRA-enhanced models significantly outperform a baseline classifier in accuracy, achieving over 80% classification performance while reducing trainable parameters. This work highlights the potential of efficient, domain-aware NLP models in capturing real-time public sentiment for financial decision-making and lays the groundwork for future integration with market forecasting tools.

1. Introduction

The modern world has created a digital connection between everyone, and this connectivity affects modern society in many ways. One of the recent and possibly most unpredictable side effects of modern connectivity is its impact on the stock market. With stock details and acquisition available to a global population via the internet, the stock market has become a game influenced by literally billions of players. This influence is then compounded further by groups of individual investors banding together via public forums to purposefully manipulate specific stock prices. This was infamously demonstrated by GameStop's stock in January 2021, where its price rose from under $5 to over $85 in just a month due to social media campaigns. As a result, investment firms must not only do the typical research needed on a company's financials but also monitor public sentiment about that company. The purpose of this project is to create a model that can extract overall sentiment from public forums such as Twitter (X) or Reddit and apply what it would mean for that company's potential investors, specifically focusing on crypto companies, as these draw some of the most attention online.

2. Background

This paper builds on recent advancements in NLP, particularly transformer-based models and parameter-efficient fine-tuning techniques. We base our work on the structure of the self-attention transformer (Vaswani, 2017) and the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), the widely used pretrained model published by Google for contextual word representations.

Recognizing the need for a domain-specific adaptation, BERTweet (Nguyen et al., 2020) was introduced as a specialized BERT variant fine-tuned on Twitter data, capturing the unique features of social media language, such as informal expressions, emoticons, references to other users, hashtags, and URLs. With sentiment analysis being a key application in NLP, the SemEval-2017 Task 4 (Rosenthal et al., 2017) provided a structured benchmark for analyzing sentiment in tweets, particularly in multi-label settings where tweets can express positive, negative, or neutral sentiments.

While transformer models have achieved remarkable success, their computational cost remains challenging for efficient deployment and task-specific adaptation. Low-Rank Adaptation (LoRA) (Hu et al., 2021) emerged as a solution to mitigate the computational expense of fine-tuning large-scale models by introducing trainable low-rank decomposition matrices. Applying LoRA to BERTweet for multi-label sentiment classification, we aim to reduce the number of trainable parameters while preserving performance on a nuanced sentiment detection in tweets.

At the same time, we review the labeling of sentiment data to make it specific to financial impact, not just general sentiment evaluation.

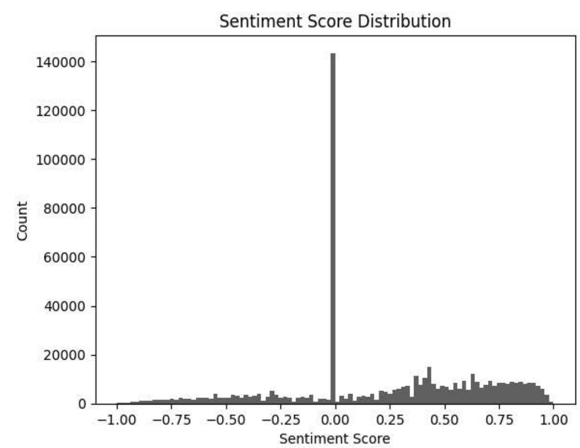
The combination of these two contributions aims to provide an efficient tool to detect trends in the dialogue in social media that can affect movements in the valuation of cryptocurrencies.

3. Methods (design and implementation)

To develop an efficient multi-label sentiment classifier for cryptocurrency-related tweets, we worked in 4 main workstreams: data preprocessing, tweet labeling, model architecture, and fine-tuning strategy.

3.1. Data Preprocessing

The data selected for this project is the dataset published by Turazzi (Turazzi, 2025) in Huggingface. This dataset contains 563,799 unique tweets from X (Twitter) associated with cryptocurrency, from Jan 2013 to Feb 2021. The labeling is done by a pre-computed sentiment score for each tweet, ranging from -1 to +1, and logits for negative, neutral, and positive sentiment. The graph below shows the distribution of the sentiment score:



Graph 1: Sentiment Score Distribution

The texts of the tweets in this dataset come in raw format, including emoticons, URLs, and mentions to other users. We leveraged the normalization work done for BERTweet (Nguyen et al.,2020) to produce a clean version of the tweets that, together with the tokenizer produced for the same model, would produce a valid input for the pretrained model.

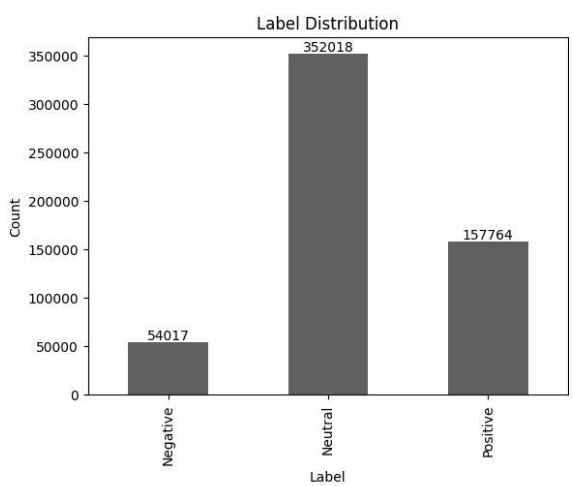
3.2. Tweet Labeling

To ensure the relevance of the data to the financial discourse surrounding cryptocurrencies (specifically buying, holding, or selling), we implemented a combination of heuristic rules and keyword selection designed to identify and exclude tweets that were deemed irrelevant, such as jokes, trolling attempts, or general non-financial content.

The second step was determining the thresholds for the sentiment score to determine positive and negative sentiment. The conditions for the threshold were to provide high precision in the labels positive and negative, while keeping the classes reasonably balanced. Through iteration, we set the thresholds at -0.05 and +0.05.

The overlap of those two criteria produced the final label:

- **Positive**: Relevant in the financial context and above the upper threshold.
- **Negative**: Relevant in the financial context and below the lower threshold.
- **Neutral**: Not relevant to the financial context or between the thresholds.



Graph 2: Final Label Distribution

3.3. Model Architecture

Our approach utilized the pre-trained BERTweet-large model as the foundation for our sentiment classifier. BERTweet-large is a large-scale Transformer-based language model that shares the same architecture as BERT_base and was pre-trained using the RoBERTa training procedure (Nguyen et al., 2020) on a corpus of 850M English tweets, making it well-suited for understanding the nuances of social media language.

We employed the AutoModelForSequenceClassification.from_pretrained function from the Hugging Face Transformers library (Wolf et al., 2020) to load this pre-trained model and attach a linear classification layer on top, configured for three output labels (positive, negative, neutral). This initial configuration served as our baseline model.

Layer (type: depth - idx)	**Param #**
RobertaForSequenceClassification	–
└─Roberta Embeddings: 2-1	–
└─Embedding: 3-1	51,471,360
└─Embedding: 3-2	526,336
└─Embedding: 3-2	1,024
└─Embedding: 3-4	2,048
└─Dropout: 3-5	–
└─RoBERTa encoder: 2-2	–
└─ModuleList: 3-6	302,309,376
Roberta Classification Head: 1-2	
└─Linear: 2-3	1,049,600
└─Dropout: 2-4	–
└─Linear: 2-5	3,075
Total params:	355,362,819
Trainable params:	**1,052,675**
Non-trainable params:	354,310,144

Table 1: Baseline Model Architecture and Parameter Summary

To enhance the model's parameter efficiency while maintaining performance, we implemented LoRA (Hu et al., 2021) and experimented with several values of alpha and rank. To implement LoRA, we used the Parameter-Efficient Fine-Tuning (PEFT) library (Mangrulkar et al., 2022) from Hugging Face. The architectural details and parameter counts for the baseline model are provided in Table 1.

Layer (type: depth - idx)	Param #
PeftModelForSequenceClassification	–
└─LoraModel:1-1	–
└─Roberta For Sequence Classification: 2-1	--
└─RobertaModel: 3-1	354,703,360
└─ModulesToSaveWrapper: 3-2	2,105,350
Total params: 356,808,710	
Trainable params: 1,445,891	
Non-trainable params: 355,362,819	

Table 2: LoRA Model Architecture and Parameter Summary

3.4. Fine-tuning Strategy

The fine-tuning process involved training both the baseline model (the added classification layer) and several LoRA-adapted versions of BERTweet-large on our labeled dataset. For the LoRA models, we explored multiple configurations of key hyperparameters, including the rank (r) of the low-rank matrices and the scaling factor (alpha). In some instances, we also adjusted the dropout rate within the model to address potential overfitting. We maintained a consistent batch size and learning rate throughout all fine-tuning experiments to ensure comparability across different configurations.

The models were trained for five epochs on a balanced dataset of 15,000 tweets, with equal samples for each of the three sentiment labels. Model performance was evaluated on a separate validation set using validation accuracy as the primary optimization metric. Additionally, we monitored the ratio between training and evaluation loss as a guardrail metric to detect overfitting during the training process.

4. Results and Discussion

After fine-tuning, the baseline model provided an accuracy of 44%. The first run of the LoRA model, with default values of rank = 8 and $\boldsymbol{\alpha}$ = 16, produced a validation accuracy of 67.8%.

After hyperparameter optimization, LoRa reached 80,7% after 5 epochs.

Following the recommendation used by Nguyen (Nguyen et al., 2020), we used the following ranges of values for LoRA parameters:

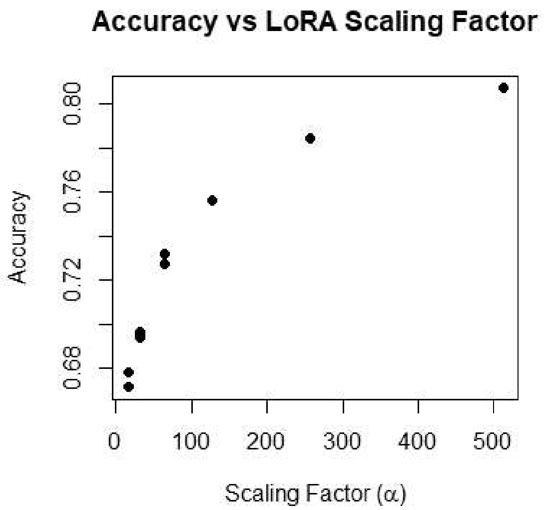
- Rank: 4,8,16
- Alpha: 16, 32, 64,128, 256, 512
- Dropout: 0.05, 0.1
- Learning Rate: 1e-5
- Batch size: 16
- Decay: 0.01

The table below shows the accuracy variation vs Rank and Scaling Factor.

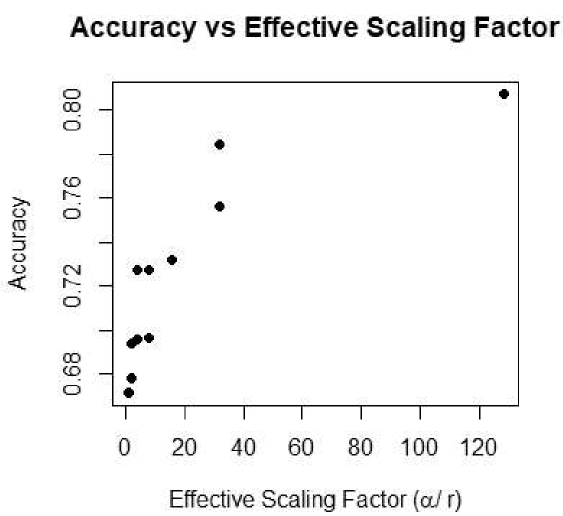
id	Rank	Alpha	Accuracy
1	8	16	0.678
2	8	32	0.696
3	8	64	0.727
4	4	32	0.696
5	4	128	0.756
6	4	64	0.732
7	16	32	0.694
8	16	16	0.671
9	16	64	0.727
10	8	256	0.785
11	**4**	**512**	**0.807***
Baseline model			0.444

Table 3: Hyperparameter optimization for LoRA. (*) The best combination found was rank = 4, Alpha = 512, and Accuracy 0.807.

Once we had specified the most effective hyperparameters, we tested the model's accuracy on a significantly larger dataset to evaluate its generalization capabilities. We created a new dataset consisting of 100,000 tweets, carefully sampled to preserve the original label distribution from the full dataset. This ensured a balanced and representative evaluation. The fine-tuned model, trained with LoRA (Low-Rank Adaptation) on a much smaller dataset of just 15,000 tweets, performed impressively on this larger test set. It achieved an overall accuracy of 75%, with particularly strong performance on class 0 (precision: 0.96) and class 2 (f1-score: 0.74). These results demonstrate the model's ability to scale well and maintain predictive power across a larger and more diverse sample.



Graph 3: Accuracy vs LoRA Scaling Factor (**α**)



Graph 2: Final Label Distribution

4.1. Baseline Comparison

The tables below show the performance differences between the baseline and the best model trained with LoRA.

Label	**Precision**	**Recall**	**F1**	**Support**
Neutral	0.72	0.56	0.63	62,437
Negative	0.18	0.40	0.25	9,581
Positive	0.38	0.42	0.40	27,982
Accuracy			**0.51**	100,000

Table 4: Out-of-sample baseline model evaluation using 100,000 tweets from the same data source.

Label	**Precision**	**Recall**	**F1**	**Support**
Neutral	0.96	0.69	0.81	62,437
Negative	0.43	0.88	0.58	9,581
Positive	0.66	0.84	0.74	27,982
Accuracy			**0.75**	100,000

Table 5: Out-of-sample optimized model evaluation using 100,000 tweets from the same data source.

5. Conclusion

This study explored the intersection of public sentiment and financial prediction, focusing on cryptocurrency discourse on Twitter (X) and leveraging transformer-based language models for multi-label sentiment classification. By combining domain-specific filtering techniques, sentiment-aware labeling, and parameter-efficient fine-tuning using LoRA, we demonstrated that it is possible to build a performant and efficient sentiment classifier tailored for the fast-paced, informal language of social media.

Our experiments revealed that LoRA not only substantially reduces the number of trainable parameters compared to traditional fine-tuning but can also outperform the baseline model in terms of classification accuracy. In particular, we achieved a significant improvement from a baseline accuracy of 44.4% to over 80% with the best LoRA configuration—highlighting the value of low-rank adaptations for practical deployment scenarios where computational efficiency matters.

Importantly, by refining sentiment classification to focus specifically on financially relevant content, this model offers a meaningful signal for analysts and investors interested in tracking crypto-related market sentiment. These findings support the notion that real-time sentiment analysis—when tailored correctly—can serve as a valuable input in financial decision-making.

Future work can expand on this by integrating time-series modeling to assess how sentiment correlates with actual market movement, or by broadening the scope to include other social platforms like Reddit. Moreover, fine-grained sentiment labels (e.g., “FOMO” vs “fear”) could be explored to capture subtler investor moods. As online dialogue continues to shape financial markets, tools like the one presented here may prove critical in anticipating and interpreting the signals behind the noise.

6. References

1. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.** 2019. 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)*. Association for Computational Linguistics.

2. **Nguyen, D. Q., Vu, T., & Nguyen, A. T.** 2020. BERTweet: A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

3. **Rosenthal, S., Farra, N., & Nakov, P.** 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics.

4. **Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W.** 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

5. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.** 2017. Attention Is All You Need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

6. **Turazzi, F.** 2025. *Cryptocurrency Tweets with Sentiment Analysis.* Kaggle. Retrieved from https://www.kaggle.com/datasets/fabioturazzi/cryptocurrency-tweets-with-sentiment-analysis.

7. **Mangrulkar, Sourab, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan**. 2022. *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning Methods*. \url{https://github.com/huggingface/peft} (Accessed April 13, 2025).

8. **Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush.** 2020. "Transformers: State-of-the-Art Natural Language Processing". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-demos.6