IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

CBITS: Crypto BERT Incorporated Trading System

GYEONGMIN KIM[1⋆], MINSUK KIM[2⋆] BYUNGCHUL KIM[3⋆] and HEUISEOK LIM[1]
[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
[2]MindsLab (AI Scientist), Gyeonggi-do 13493, South Korea
[3]HighDev (Chief Technology Officer), Synotex (Chief Data Officer), Busan 47291, South Korea

Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)
These authors contributed equally to this work(⋆).

ABSTRACT Most textual analysis-based trading approaches in cryptocurrency (crypto) involve lexical, rule-based methods for extracting news sentiments. Furthermore, general purpose language models (LMs) are not always suitable for the crypto domain due to jargons that are not covered in general purpose texts. This study answers the question of "Is it possible that the LMs can profit by effectively applying the sentiment score of the NLP task with chart score in the BTC trading system?" by focusing on the effectiveness of both scores, which significantly affect the profit of the trading system. We introduce **CBITS**: **C**ryptocurrency **B**ERT **I**ncorporated **T**rading **S**ystem based on pre-trained LMs for Korean crypto sentiment analysis to aid Bitcoin (BTC) trading models. We specifically pre-trained crypto-specific LMs, which are transformer encoder-based architectures. Along with our pre-trained LMs, we also present our custom fine-tuning dataset used to train our LMs on the BTC sentiment classification task and show that using sentiment scores along with BTC chart data boosts the performance of BTC trading models and allows us to create a market neutral trading strategy.

INDEX TERMS Cryptocurrency, Korean Pre-trained Language Model, Sentiment analysis, Bitcoin trading models, Korean Pre-trained Language Model

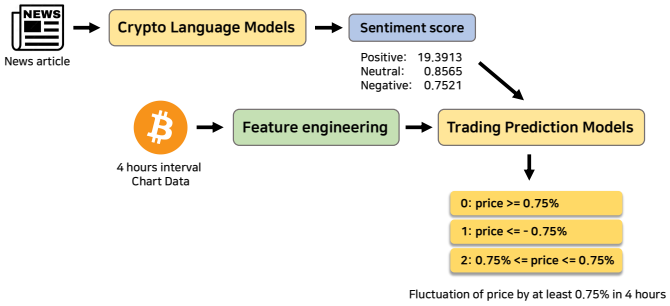
I. INTRODUCTION

Since the advent of the cryptocurrency (crypto) market that is now a trillion dollar market as of 2022, Bitcoin (BTC) [25] has attracted attention in the research community due to its unconventional volatile price fluctuations and unpredictability. To address the challenge of predicting BTC price movements, numerous works were published. It is well known in the domain of finance that news sentiments are helpful in predicting the price fluctuations of financial assets [13, 23]. In order to effectively capture crucial trade signals from news articles in an automated manner, This begs the question of "Why there are not any publicly available language models (LM) for the crypto and the blockchain field?". Since the emergence of BTC, the crypto market grew quickly and there were many financial investors trying to gain an advantage in the market using statistical [16] as well as machine learning [4] methodologies. Especially, methods incorporating textual analysis [1, 27] for price predictions of crypto began to emerge.

Unlike the financial field, which has a domain adapted pre-trained LM, the textual analysis techniques used in this crypto domain mainly seemed to be achieved through lexical and rule based methods. Following the previous success in financial textual analysis achieved by FinBERT [2], we propose **CBITS**: **C**ryptocurrency **B**ERT **I**ncorporated **T**rading **S**ystem based on pre-trained LMs that are applicable to the crypto and the blockchain domain in Korean. In this research, we focus on calculating crypto news sentiment scores and use these calculated scores to enhance the performance of our BTC trading model.

Korean language which is agglutinative in its morphology is challenging due to the intermediate characteristics positioned between isolating and inflectional language. Basically, the word in Korean language is composed of both eojeol and grammatical morpheme. The first challenge is that the meaning of an eojeol can vary depending on the grammatical morpheme located behind the eojeol. For example, the meaning of the noun '그 (He)' can change depending on the josa postpended to it, such as '그는 (He is)' or '그에게 (to him)'. The second challenge is that a jamo which is consisted of a consonant and vowel, can be represented differently depending on the meaning of the morpheme. For

FIGURE 1: An overview of **CBITS** with sentiment analysis based on crypto LMs.



example, '새(bird)' composed of 'ㅅ' and 'ㅐ' is transformed in meaning to '소' which means 'cow' due to the variations in vowel. For these reasons, domain-specific LM should be considered differently than general LM, and we designed pre-trained LM with Sentencepiece [17], which is a language-independent subword tokenization algorithm that does not require language-specific processing. We collected various Korean crypto dataset to pre-train three transformer based LMs: BERT [11], RoBERTa [22] and DeBERTa [14]. We also created our custom fine-tuning dataset to train our pre-trained LMs. Our fine-tuning corpus is designed to let our **CBITS** learn how the news article may affect the prices of BTC.

The sentiment classification task is not simply calculating the polarity of the news articles, but is focused on how they may affect the movement of BTC prices. Figure 1 is an overview of our **CBITS** with scores both of sentiment polarity and chart in the BTC trading system. We design a trading model for the BTC Tether (USDT) perpetual derivatives market motivated by the classification approach described in [34], and then supplemented this trading bot with news sentiment features from our fine-tuned LMs to show that news sentiment scores from pre-trained crypto LMs boost the performance of this trading bot. We also show that TabNet [3] is an optimal choice for the BTC trading model by experimentally verifying that it produces the most robust results among our candidate models. Our trading models were purposely designed to trade in the derivatives market as we wanted to create a market-neutral strategy (i.e., the model can profit in both the bullish and the bearish markets).

The contributions are summarized as follows:

1) We propose a novel Korean crypto news sentiment dataset specifically tailored for BTC. News is labeled positive, negative, or neutral depending on the effect of BTC price movements.
2) We design Korean crypto domain specific pre-trained LMs that are fine-tuned on our dataset.
3) We propose classification based trading models for trading BTC in the derivatives market (BTC/USDT Perpetual).
4) We prove the effectiveness of our **CBITS** enhancing the performance of our BTC trading models.

II. RELATED WORK

Previous studies have been incorporating deep learning approaches for textual analysis and a notable example is the state augmented reinforcement learning (RL) framework in finance field [24, 33]. The state augmentation is achieved by incorporating news sentiment into an RL framework by training a hierarchical attention network [32] with three different word embedding techniques, and they proved that this framework beats the buy and hold as well as other online portfolio selection methods [20].

Recently, Mohan et al. [24] predict stock prices by feeding an LSTM network [15] with data both of stock price and news headline sentiment text from the fine-tuned FinBERT. They showed that using the textual information significantly boosted the performance of intraday stock trading models. There were also studies using other approaches of incorporating sentiment features such as Sonkiya et al. [28], which leverages news sentiments from BERT and feeds these textual features to Generative Adversarial Networks [12], as well as Chen et el. [7], which makes use of contextualized embeddings generated from news headlines for price prediction. Even though other recent approaches include more sophisticated deep learning architectures for crypto textual polarity classification [27, 30], none of them involve the use of pre-trained LMs adapted to the crypto domain.

III. DATA ANNOTATION PROCESS

To improve the quality of annotations, we conduct the annotation process based on expertise in the crypto domain.

A. CRYPTO NEWS CORPUS COLLECTION

For the fine-tuning LMs of our **CBITS**, we crawled the data from a mainstream Korean crypto news source website called Coinness Korea (CK) [1], from 2018-01-19 14:00 UTC to 2022-04-16 00:00 UTC. The CK is dedicated to delivering crypto and blockchain trade information day and night to help actors make better crypto investment decisions.

1) Annotation Process

The dataset was labeled by a professional day trader and was peer reviewed by another cross-checking day trader. During the labeling process, the annotators were provided with sentiment classification results from FinBERT as well as the information of bear/bull market votes from the CK website for reference. Furthermore, the news data was not analyzed in isolation, but the annotators were instructed to observe both the chart movements and the news together for labeling. Along with the additional information, the annotators adhered to the annotation guidelines when tagging news

[1] https://coinness.live/

sentiment labels. Details of the guidelines are described in Section III-A2.

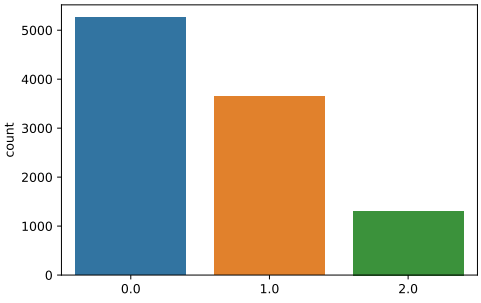
2) Annotation Guidelines

- Because our trading bot will trade BTC in the derivatives market, the news sentiments are labeled according to how this particular news may affect BTC prices. So if the news article talks about a crypto coin that is deemed to not affect BTC whatsoever, even if the general sentiment of that news is positive or negative, that news is labeled neutral. In other words, the sentiments are focused on BTC.
- The price increase, bullish news, and investment attraction of large crypto coins up to the top 15th in market capitalization are also reflected as good news for BTC and given a positive label. However, if BTC dominance dropped sharply at that point, a negative label is assigned instead.
- News containing information that large amounts of BTC are deposited to exchange is given a negative label because it is usually the case that many people are willing to either sell or short BTC. However, if the deposit amount is small or seems like an inflow of stablecoins, we give it a neutral label.
- Transfer of crypto from large exchanges (e.g. Binance, Bybit) to an anonymous wallet or a transfer of crypto from anonymous to anonymous is labeled as neutral.
- Most of Elon musk's provocative tweets are labeled as being positive, as Elon has never led a decline in prices, but we believe his influence over the market is gradually weakening over time.
- Aside from Elon, opinions from other influencers or public figures are labeled as neutral. However, words from the US senate or the fed officials are deemed important and are labeled either positive or negative.
- News about a company or a country using or investing in a crypto is generally labeled as being positive. However, if the company is not well known or deemed insignificant then we give the news a neutral label.
- All breaking news that talks about the number of times BTC was mentioned/tweeted is carefully analyzed with the chart data and is given either a label of positive or negative. If there are lots of Twitter mentions for particular crypto, it is usually an indication that there was a large fluctuation for this crypto.
- Rising price index, rising oil prices, rising gold prices, and rising government bonds are major causes of the NASDAQ decline. Since BTC and NASDAQ tend to be coupled with each other we give a negative label to such news articles.
- News about small investment attractions is labeled neutral, whereas large fund and institutional investment attractions are given positive labels.
- Any small-scale investment attraction based on large coins (top 15 market cap) is given a neutral label.
- News that mentions that NFTs or metaverse platforms reached high revenue or trading volume is given a positive label.
- Any indication of decreasing BTC dominance is given a negative label.
- News related to the introduction or use of BTC in any country in the world is in general given positive labels.
- Decrease in BTC hash rate is considered negative, whereas an increase in hash rate is considered positive.
- News about a country's economic collapse is considered negative.
- The lockup release of large cryptocurrencies is in general considered neutral.
- News about token burns of cryptocurrencies with a large market cap is considered positive in general.

B. SILVER DATASET

Annotated gold dataset is described in Figure 2. Label 0 is positive, label 1 is negative and label 2 is neutral. Approximately 77% of the labels are neutral, 15% of the labels are positive and 8% of the labels are negative. Due to the imbalance of sentiment labels in the annotated gold dataset, we leveraged the best fine-tuned LM described in Section V-B to extract only the positive and the negative news among our unlabeled news data from 2018-01-19 to 2022-04-16. The **CBITS** labeled 18,005 news corpus as either positive or negative, and this silver dataset was also cross-checked by the annotators. Once the true positives and negatives were selected from the silver dataset, they were combined with the gold dataset to be used for statistical testing in Section IV-A.

FIGURE 2: Sentiment label distribution



C. CHART DATA

We collect 4-hour interval BTC/USDT chart trading data from Binance [2], which is a large exchange in the crypto domain. After collecting and feature engineering the chart data, we ended up with data from 2017-08-23 16:00 UTC to 2022-04-23 00:00 UTC. In total, we have 35 chart related input features excluding the target variable. The training/validating/testing split was performed that the training data ranges from 2017-08-23 16:00 UTC to 2022-01-27

[2]https://www.binance.com

16:00 UTC, validating data ranges from 2022-01-28 00:00 UTC to 2022-02-13 20:00 UTC, and testing data ranges from 2022-02-14 00:00 UTC to 2022-04-23 00:00 UTC. The time ranges do not overlap to ensure that no look-ahead bias exists.

D. CHART FEATURES USED FOR BTC TRADING MODELS

In this section, we describe the chart features in detail. Instead of using raw chart data (e.g., high, low, open, close, volume), we conduct feature engineering to extract useful trading information. Feature engineering when it comes to chart data can become extensive as there are hundreds of existing technical indicators out there. Since the size of the chart data is relatively small and adding too many features may hinder the model's performance, we picked only a few technical indicators and chart features. Most of these features are related to volatility or momentum as we considered that capturing changes in buying or selling momentum. It is important for predicting either long or short positions at the current timestamp.

The description of features we used is as follows:

- **Differencing**: This is simply the ratio of raw chart features across different time periods. Since we are using 4-hours interval chart data, the difference between a single time period (e.g. t and $t-1$) would be 4-hours. The first differencing of close prices would be calculated as follows:

$$\text{First Difference of Close} = \frac{\text{Close}_t}{\text{Close}_{t-1}}$$

  In general, the K_{th} differencing of the close prices is simply:

$$K_{th} \text{ Difference of Close} = \frac{\text{Close}_t}{\text{Close}_{t-K}}$$

  We carried out this differencing procedure for all open, high, low, close, and volume and used $K = 1, 2, 3, 4, 5$ for differencing.
- **EBSW**: The even better sinewave (EBSW) is a variation of the Hilbert sine wave and it is an indicator that can possibly inform the model about the bullish and bearish cycle of prices. Typically, an EBSW value greater than 0.85 suggests that the asset is overbought whereas a value smaller than -0.85 suggests that the the asset is oversold. The pandas-ta[3] library was used for EBSW calculation.
- **Chaikin Money Flow**: The chaikin money flow (CMF) is an indicator that is used to monitor both accumulation and distribution of an asset over a specified period of time. The value range is from -1 to +1 and any crosses above or below 0 can be used to identify certain momentums in buying and selling. The pandas-ta library was used for CMF calculation. The default period of 20 is used for CMF.

[3]https://github.com/twopirllc/pandas-ta

- **VWAP/Open**: Ratio of volume weighted average price (VWAP) and open price. The calculation of VWAP first involves calculating the typical price TP, which is simply the average of low, high and close prices.

$$\text{TP} = \frac{\text{Low} + \text{High} + \text{Close}}{3}$$

  VWAP at time t is calculated by the following formula:

$$\text{VWAP} = \frac{\sum_{i=1}^{t} \text{TP}_i \times \text{Volume}_i}{\sum_{i=1}^{t} \text{Volume}_i}$$

- **Other ratio features**: Ratios from raw chart features such as High/Low, High/Open, Close/Open, and Low/Open were used to provide price volatility information to the model.
- **Time Information**: We use the hours, days, and months information. We deliberately excluded the years' information as years (such as 2020, 2021, 2022, etc) do not follow a cyclical pattern like hours, days, and months.

IV. METHODOLOGY

In Section IV-A, we describe the strong correlation between news sentiments and BTC returns. Motivated by this fact, we present a classification based approach for our BTC trading model in Section IV-B.

A. SENTIMENT SCORE AND BTC RETURNS

As it is a pervasive problem in trading research whereby some features are fed to a machine learning algorithm, and the modification of the structure is repeated until the desired backtest result shows up, we perform Kolmogorov-Smirnov (KS) test to first show that the BTC's price changes right after a certain sentiment score are in fact differently distributed than the price changes right after a neutral or no sentiment scores, proving the statistical significance of the correlation between news sentiment scores and BTC's price changes [9, 10].

Here, we specify sample A_s and B where sample A_s is a set of BTC's price changes right after the sentiment score s, in which $s = 0$ represents positive, $s = 1$ represents negative, $s = 2$ represents neutral sentiment score according to the labels in Figure 2. Sample B shows a set of price changes right after neutral or no sentiment score. This assumption follows intuitive reasoning as if the news sentiment scores have zero impact to the BTC's price changes, then the distribution of sample A_s will be similar to sample B, i.e., the price changes right after the said news sentiment scores will be no different from when there were neutral or no sentiment scores. Applying the KS test, we show that the probability of observing sample A and sample B given that they are sampled from the same distribution is statistically low.

To first find the KS test statistic of sample A and sample B, we find their respective empirical cumulative distribution function (ECDF) by binning each sample and finding the empirical number of values less than or equal to a certain value.

TABLE 1: Probability of observing the KS test statistic of each data frequency with positive and negative news given the assumption that samples A and B are of the same distribution.

Data Frequency	P-Value (s=0) (%)	P-Value (s=1) (%)
1-hour	0.0152	**0.0000**
2-hours	0.0266	0.0011
3-hours	0.0206	0.0018
4-hours	**0.0112**	**0.0000**

$$F_n(a) = \frac{\text{number of sample's elements} \leq a}{n} \quad (1)$$

where n represents the total number of the size of the sample. To apply the ECDF, we first have to find appropriate `bins` where the sample's element can be "binned" for calculating the ECDF in practice. Although the size of the `bin` can arbitrarily be small, we apply Freedman-Diaconis rule, which is used for performing a histogram as a density estimator, as a rule of thumb for finding the size of the `bin`. Given discrete empirical measurements, this rule used to find the `bin` width is defined as:

$$\texttt{bin width} = 2.0 \times \text{IQR}(x) \times n^{-1/3} \quad (2)$$

where IQR(x) represents interquartile range of the data and n represents the number of elements in the data. The number of `bins` is as follows:

$$number\ of\ \texttt{bins} = \frac{max(sample) - min(sample)}{\texttt{bin width}} \quad (3)$$

Upon finding the appropriate size of the `bin`, we iterate through each `bin` and find the number of examples less than or equal to the `bin`'s upper value to find the ECDF for both sample A and sample B, named $\text{ECDF}_{A,b}$ and $\text{ECDF}_{B,b}$ respectively where $\text{ECDF}_{A,5}$ represents sample A's ECDF value at 5_{th} `bin`. The KS test is based on the intuitive idea that if two samples are sampled from the same distribution, then their ECDF's must be identical, and thus the "distance" between these two ECDF's must be small, with enough data making the distance arbitrarily close to zero. The KS test statistic for two empirical cumulative distribution functions are as follows:

$$\text{KS} = \sup|\text{ECDF}_{\text{A, b}} - \text{ECDF}_{\text{B, b}}|_{\text{for all b}} \quad (4)$$

We then apply a permutation test where sample A and sample B are combined and permutated. It randomly split to create new sets of two samples sample A′ and sample B′, and then perform monte carlo simulation to find the estimated distribution of the "distance" between newly randomly sampled samples A′ and B′. The idea behind the permutation test is that if sample A and sample B are in fact sampled from the same distribution, then their initially found ECDF distance will be expected to be similar to permutated and newly sampled sample A′ and sample B′. Based on this simulation, we can then calculate the p-value given that those sample A and sample B are in fact sampled from the same distribution.

Before directly applying KS test on sample A and sample B, we define the range of `t` where `t = 1` represent hourly prices, and `t = 2` represent every other hour's prices. The reason for conducting KS test not only for different sentiment scores (positive or negative) but also for the range of `t` is that financial prices such as BTC's prices are non-stationary, and have low signal-to-noise ratio[29]. Thus, formulating a trading strategy that is based on a small value of `t` (e.g., t = 1 minute), then the noise in the data might overwhelm the signal in the time series, and therefore overwhelm the predictive impact that the crypto news have on BTC's returns. we perform and present the probability of observing the KS test statistic for each data frequency `t` and for each sentiment score $s = 0$ (positive), and $s = 1$ (negative) in Table 1.

B. BTC TRADING SYSTEM

1) Classification Based BTC Trading Model

We tackle the problem of BTC trading by leveraging a classification approach. Our target variable is defined by three classes. If the equation is as follows:

$$u_{t+1} = \frac{\text{high}_{t+1} - \text{close}_t}{\text{close}_t}, \quad v_{t+1} = \frac{\text{low}_{t+1} - \text{close}_t}{\text{close}_t} \quad (5)$$

The labels are defined as follows:

- c_0: BTC price rises by at least 0.75% within the next 4-hours i.e., $u_{t+1} >= 0.0075$
- c_1: BTC price drops by at least 0.75% within the next 4-hours i.e., $v_{t+1} <= -0.0075$
- c_2: BTC price change within the next 4-hours is less than 0.75% i.e., $u_{t+1} < 0.0075$ and $v_{t+1} > -0.0075$

We defined the threshold at 0.0075 to account for the trade commissions and bid-ask spread in practice when both $u_{t+1} >= 0.0075$ and $v_{t+1} <= -0.0075$ happen at the same time. We inspected that this case happens in approximately 14.5% of the data. In a situation when both the short and long position results in at least a 0.75% profit, we favored long over short.

Class distributions are 51.5% for c_0, 35.7% for c_1 and 12.8% for c_2. When both c_0 and c_1 occur we favoured c_0 which explains why most of the labels are c_0. Class c_2 has the least occurrence which suggests that it is quite common for BTC to fluctuate by at least 0.75% (in either direction) within the next 4 hours.

2) Comparison of chart only models

We considered three candidate models to be used for our classification based trading bot.

TABLE 2: Crypto based pre-trained LMs: BERT, DeBERTa, RoBERTa. The hyperparameters in the middle refer to #ML = max sequence length, #L = learning rate, #U = the number of total updates, and #P = perplexity. The right side shows the results of fine-tuned LMs' score using crypto news sentiment corpus, which is annotated by professional day trader in Section III-A1.

PLM	Hyperparameters				Test Set F1-score (%)
	#ML	#L	#U	#P	Crypto News Sentiment
CryptoBERT	512	1e-3	200K	10.26	81.42
CryptoDeBERTa	512	1e-3	200K	8.12	81.75
CryptoRoBERTa	512	1e-6	200K	4.68	**83.01**
mBERT					79.83
XLM-RoBERTa					80.63

- **LSTM**: The long short term memory [15] network is a type of recurrent neural network (RNN) that aims to solve the long-term dependency problem of RNN via gating mechanisms.
- **XGBoost**: eXtreme Gradient Boosting [8] is a scalable and highly accurate implementation of gradient boosting.
- **TabNet**: TabNet [3] is attention based neural network architecture specifically designed to tackle problems involving tabular data. It uses soft feature selection to focus on features that are important and this process is accomplished via a sequential multi-step decision mechanism.

We purposefully did not include well known attention based time series forecasting models such as the temporal fusion transformer [21] and the informer [35] since we are dealing with tabular data and the dataset size is not large enough to train complicated neural network architectures. To account for the class imbalance, we use balanced class weights when training all three models.

V. EXPERIMENT

A. EXPERIMENTAL PRELIMINARIES

a: Optimal pre-trained LMs for sentiment classification

We proceed with a practical preliminary step to prove the optimal fine-tuned LM for the sentiment classification task. As shown in Table 2, we pre-trained three crypto LMs: BERT, RoBERTa, DeBERTa. Perplexity (PPL) corresponding to the intrinsic evaluation represents the degree of confusion of the model and is primarily used as an evaluation metric for LMs. We also pre-experiment with the effectiveness of sentiment classification by fine-tuning the three crypto pre-trained LMs with two original multilingual LMs: mBERT and XLM-RoBERTa.

As a result, the ideal LMs compared with the five LMs is CryptoRoBERTa, which shows the best performance on the crypto news sentiments analysis task while showing a low PPL. In other words, it can be concluded that this **CBITS** is an optimal model that effectively calculates news sentiment scores used to aid our BTC trading models.

b: Pre-training corpus

To design LMs optimized for cryptocurrecy domain, data used for pre-training **CBITS** were mostly Korean crypto and blockchain related news. It was intended because news articles follow a format that is easy to work with, and also covers the most recent information as well as technical terms used in the crypto domain. Aside from news sources, texts from crypto blogs, crypto mining community forums, and Wikipedia were used. Overall, we managed to collect 900K texts that were amassed to 880 MB in size for pre-training **CBITS**.

c: Experimental setup

For our **CBITS**, we pre-trained the model using huggingface [31] for BERT and DeBERTa with 2 A100 GPUs, and FAIRSEQ[4] [26] for RoBERTa with 4 A6000 GPUs.

For the crypto sentiment classification task, we set consistent experimental environment following fine-tuning hyperparameters: *learning rate* = $1e^{-5}, 3e^{-5}, 5e^{-5}$; *batch size* = 32; and *max epoch* = 10 except when early stopping occurs. Considering that the performance of the model may vary depending on the initialization value, the average score with three learning rate and five random initialization values is recorded.

B. FINE-TUNING CRYPTO SENTIMENT ANALYSIS TASK

The pre-trained crypto LMs were subsequently fine-tuned on our crypto news dataset. To enhance our experimental procedure, we conducted an evaluation with ten-fold cross-validation, i.e., 10 times for each model, where 8:1:1 for training, validating and testing, respectively. To verify the effectiveness of our **CBITS**, two representative multilingual LMs were also fine-tuned and our model recorded about 3 points higher. The detailed scores are on the right side of Table 2.

C. INCORPORATING NEWS SENTIMENT SCORES

Given the performance of chart only models in Section IV-B2, we experimented with adding news sentiment scores from our current best fine-tuned crypto LM, CryptoRoBERTa. As we are dealing with a 4-hours interval trad-

[4]https://github.com/pytorch/fairseq

TABLE 3: Performance of chart only models & chart + news sentiment score models. The highest score is showed in bold text, and the underline indicate the highest gap of score when using the crypto LMs.

Model	Accuracy (%)	F1 score (%)
LSTM (+ CryptoRoBERTa)	68.46 → 69.83 (+1.37)	65.48 → 66.68 (+1.20)
XGBoost (+ CryptoRoBERTa)	**70.66** → **72.62** (+1.96)	64.35 → 65.74 (+1.39)
TabNet (+ CryptoRoBERTa)	69.19 → 72.37 (+3.18)	**66.45** → **68.36** (+1.91)

TABLE 4: Performance of chart + news sentiment score models with top 5, 10 semantic search method.

Model	Accuracy (%)		F1 score (%)	
Top K	Top 5	Top 10	Top 5	Top 10
LSTM (+CryptoRoBERTa)	70.41 (+0.58)	70.42 (+0.59)	67.30 (+0.62)	67.31 (+0.63)
XGBoost (+CryptoRoBERTa)	73.34 (+**0.72**)	72.37 (-0.25)	67.27 (+**1.53**)	65.81 (+0.07)
TabNet (+CryptoRoBERTa)	70.66 (-1.71)	73.35 (+**0.98**)	67.91 (-0.45)	69.83 (+**1.47**)

ing bot, we simply collected all the CK news that were uploaded within the 4-hours window, calculated the sentiment scores for each of them and added these sentiment scores. The added sentiment score is fed into the model along with our feature engineered chart data to output a prediction. We only used positive and negative sentiment scores as we consider neutral news as noisy data. After grouping our CK news data into 4-hour intervals, approximately 62.56% of the chart data had news available within that 4-hour time intervals. For intervals that had no news data at all, zero positive and negative sentiment scores were given as inputs.

The results of adding news sentiment scores are shown in Table 3. All models show performance improvements of +1.37%, +1.96%, +3.18% on accuracy score, and +1.20%, +1.39%, +1.91% on F1 score, respectively compared to chart only models. This is probably due to the fact that XGBoost and TabNet are more suited for problems involving tabular data. XGBoost has the highest accuracy, but TabNet has the highest F1 score, suggesting that TabNet is the more robust model.

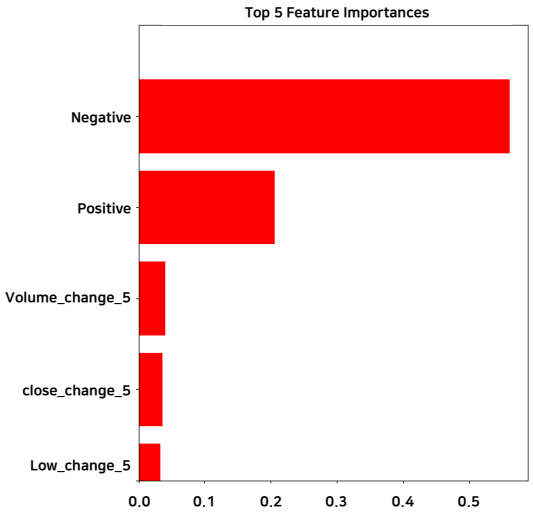
D. SEMANTIC SEARCH BASED SENTIMENT SCORE CALCULATION

We observed that some news articles exceed the 512 token length limitations of LMs. This motivated us to use semantic search to find the most relevant top K sentences from the content to the title of the news article, where $K = 5, 10$.

The model used for calculating semantic similarity was KR-SBERT [5], which is a siamese network consisting of a pre-trained KR-BERT fine-tuned on the KLUE NLI dataset and augmented with KorSTS dataset. The top 5 and top 10 most relevant sentences calculated by the cosine similarity between the title embedding and the content embedding were used instead of the entire content for calculating the news sentiment scores. Models trained using these methods are referred to as RoBERTa top 5 and RoBERTa top 10. The results are shown in Table 4.

[5]https://github.com/snunlp/KR-SBERT

FIGURE 3: Top 5 most important features



We observe that the best accuracy and F1 score are achieved by TabNet RoBERTa top 10. In general, the F1 score for all three models improves when using the top 10 semantically similar sentences from the content. Because TabNet shows the most robust results with the best accuracy and F1 score, we decided to use TabNet as the trading model going forward. Observing the feature importance plot from TabNet RoBERTa top 10 in Figure 3, we show that the most weights are given to the negative and the positive sentiment scores calculated by our **CBITS**, followed by a volume related feature. This suggests that sentiment scores played an essential role in the trading model.

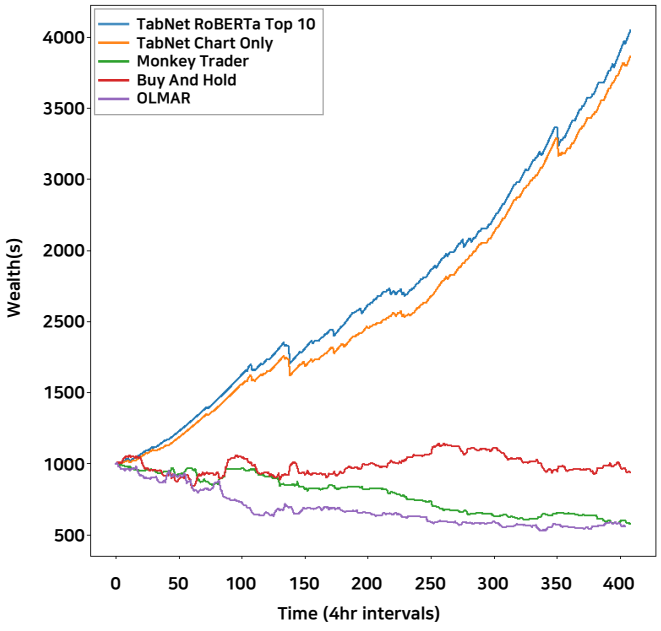
E. BACKTEST RESULTS

To verify the performance of the trading model, we conducted a backtest of our TabNet models on the test dataset. Along with these, we also considered some other baseline trading methods for comparison.

TABLE 5: Backtest results of various trading models.

Models	Profit (%)	Volatility	MDD (%)
TabNet RoBERTa Top10	**304.65**	0.0062	**-7.86**
TabNet (chart only)	286.51	**0.0060**	**-7.86**
BAH	-5.92	0.0133	-21.57
Monkey Trader	-46.89	0.0110	-48.63
OLMAR	-44.00	0.0133	-47.58

FIGURE 4: Backtest results of various trading models.



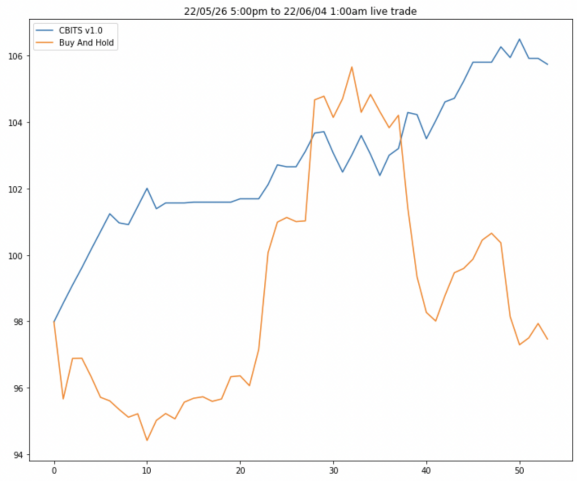
- **BAH**: Buy and Hold strategy. It simply buys BTC and holds it during the entire testing period.
- **Monkey Trader**: A random agent, that chooses long, short or hold (doing nothing) with uniform probability.
- **OLMAR**: Online Portfolio Selection with Moving Average Reversion (OLMAR) strategy. OLMAR [19] assigns portfolio weights to each of the assets in the portfolio. In our case, the two assets will be the cash agent and BTC. When OLMAR assigns at least 60% of the weight on BTC, then we take this as a signal to take a long position, and if OLMAR places at least 60% of the weight on the cash agent, we take this as a signal to take a short position. Otherwise, we resort to 'hold'.

The TabNet models predict what action to take (long, short, hold) and their profits are recorded after every action they make. When trading with the TabNet models, we assume a 0.75% take profit and no stop loss. No take profits or stop losses are assumed for other models. Furthermore, commissions of 0.04% are used to simulate trading as a market taker in the Binance USDT market. When actually calculating the wealth achieved for the backtest we used 0.08% as the commissions (twice the actual commissions) in order to account for factors such as the bid-ask-spread and the slippage that are difficult to simulate for backtesting. The initial seed money for all the models is set to $1000.

Figure 4 shows the backtest results (portfolio value) of various BTC trading models and Table 5 shows their total percentage profits. Our TabNet models outperform other trading methodologies by a wide margin, with the best profit of 304.65% achieved by the TabNet RoBERTa top 10. All the other trading models result in a negative profit during the testing period.

VI. LIVE RUN RESULTS & LIMITATIONS

FIGURE 5: Backtest on Live Run



In Figure 5, We carried out live run experiments using our proposed **CBITS** architecture from 2022/05/26 5:00pm to 22/06/04 1:00am. Our live run was conducted on Bybit's BTCUSDT perpetual market, with taking profit set at 0.75% and stop loss naively set at 2.0%. The initial seed was approximately $100. We compared the performance of **CBITS** with the performance of the buy and hold strategy, and **CBITS** managed to achieve a final profit of +7.926%, a maximum profit of +8.700%, and a maximum drawdown of -1.274%. On the other hand, buy and hold resulted in a final profit of -0.527%, maximum profit of +11.927%, and maximum drawdown of -7.926%. During this testing period, **CBITS** outperformed buy and hold by a large margin and showed lower risk than buy and hold. By observing these results, we also show the limitations in the current **CBITS** methodology.

Since the take profit is set at 0.75%, in extremely bullish markets, **CBITS** cannot achieve more profit than buy and hold. Similarly, in extremely bearish markets, our **CBITS** cannot achieve more profit than a strategy that shorts during that period. A more intricate take profit or stop loss strategy can potentially take our proposed **CBITS** to the next level.

VII. CONCLUSION AND FUTURE WORK

In this research, we created LM adapted to the crypto domain to calculate crypto news sentiments and showed that these sentiment scores help improve the performance of BTC trading models. Our primary focus in this paper was proving the effectiveness of news sentiments, but one might be able to further improve the performance of the trading bot by considering other features along with news sentiments, such as on-chain data [6], BTC dominance [18]. If more news and chart data are collected, we may be able to experiment with more complicated multi-modal structures such as work done by [5]. Crypto LMs may also be applied to other NLP tasks such as detecting pump-and-dump scheme articles, crypto named entity recognition, and crypto news clustering based on similarity analysis.

REFERENCES

[1] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra. Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3): 1, 2018.

[2] D. Araci. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019.

[3] S. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8):6679–6687, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16826.

[4] A. Barnwal, H. P. Bharti, A. Ali, and V. Singh. Stacking with neural network for cryptocurrency investment. In 2019 New York Scientific Data Summit (NYSDS), pages 1–5, 2019. .

[5] Z. Boukhers, A. Bouabdallah, M. Lohr, and J. Jürjens. Ensemble and multimodal approach for forecasting cryptocurrency price, 2022. URL https://arxiv.org/abs/2202.08967.

[6] I. Chalkiadakis, A. Zaremba, G. W. Peters, and M. J. Chantler. On-chain analytics for sentiment-driven statistical causality in cryptocurrencies. Blockchain: Research and Applications, 3(2):100063, 2022. ISSN 2096-7209. . URL https://www.sciencedirect.com/science/article/pii/S2096720922000033.

[7] Q. Chen. Stock movement prediction with financial news using contextualized embedding from bert, 2021. URL https://arxiv.org/abs/2107.08721.

[8] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. . URL http://doi.acm.org/10.1145/2939672.2939785.

[9] T. Chordia, A. Goyal, and A. Saretto. p-Hacking: Evidence from Two Million Trading Strategies. Swiss Finance Institute Research Paper Series 17-37, Swiss Finance Institute, Aug. 2017. URL https://ideas.repec.org/p/chf/rpseri/rp1737.html.

[10] M. L. De Prado. The 10 reasons most machine learning funds fail. The Journal of Portfolio Management, 44(6): 120–133, 2018.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. . URL https://www.aclweb.org/anthology/N19-1423.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[13] L. Guo, F. Shi, and J. Tu. Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science, 2(3):153–170, 2016.

[14] P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[16] Z. Kakushadze. Cryptoasset factor models. Algorithmic Finance, 7(3-4):87–104, 2018.

[17] T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. . URL https://aclanthology.org/D18-2012.

[18] A. Kulal. Followness of altcoins in the dominance of bitcoin: A phase analysis. Macro Management Public Policies, 3(3), 2021. ISSN 2661-3360. . URL https://ojs.bilpublishing.com/index.php/mmpp/article/view/3589.

[19] B. Li and S. C. H. Hoi. On-line portfolio selection with moving average reversion. In Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML'12, page 563–570, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

[20] B. Li, D. Sahoo, and S. C. Hoi. Olps: A toolbox

for on-line portfolio selection. Journal of Machine Learning Research, 17(35):1–5, 2016. URL http://jmlr.org/papers/v17/15-317.html.

[21] B. Lim, S. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4):1748–1764, 2021. ISSN 0169-2070. . URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

[22] 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, 2020. URL https://openreview.net/forum?id=SyxS0T4tvS.

[23] T. Loughran and B. McDonald. Textual analysis in finance. Annual Review of Financial Economics, 12: 357–375, 2020.

[24] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu. Stock price prediction using news sentiment analysis. In 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), pages 205–208, 2019. .

[25] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, page 21260, 2008.

[26] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. . URL https://aclanthology.org/N19-4009.

[27] A. P. Pawlicka Maule and K. Johnson. Cryptocurrency day trading and framing prediction in microblog discourse. In Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 82–92, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. . URL https://aclanthology.org/2021.econlp-1.11.

[28] P. Sonkiya, V. Bajpai, and A. Bansal. Stock price prediction using bert and gan. arXiv preprint arXiv:2107.09055, 2021.

[29] F. J. P. Thomas Dimpfl. Nothing but noise? price discovery between cryptocurrency exchanges. 1(19): 34, 2019.

[30] A.-D. Vo, Q.-P. Nguyen, and C.-Y. Ock. Sentiment analysis of news for effective cryptocurrency price prediction. International Journal of Knowledge Engineering, 5(2):47–52, 2019.

[31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. CoRR, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. . URL https://aclanthology.org/N16-1174.

[33] Y. Ye, H. Pei, B. Wang, P.-Y. Chen, Y. Zhu, J. Xiao, and B. Li. Reinforcement-learning based portfolio management with augmented asset movement prediction states. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 1112–1119, 2020.

[34] D. Zhao, A. Rinaldo, and C. Brookins. Cryptocurrency price prediction and trading strategies using support vector machines. arXiv: Trading and Market Microstructure, 2019.

[35] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35 (12):11106–11115, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

GYEONGMIN KIM received the B.S. degree in computer science and information security from Baekseok University, Cheonan, South Korea, in 2017. He is currently pursuing the Ph.D. degree in computer science and engineering with Korea University, Seoul, South Korea. Since 2017, he has been a Researcher with the Natural Language Processing and Artificial Intelligence Laboratory, Korea University. His research interests include natural language processing, multimodal learning, machine reading comprehension with neural symbolic knowledge. Particularly, his research focuses on how machines can understand like humans.

MINSUK KIM received his B.S degree in Mathematics from Stanford University, Stanford, CA, United States in 2021. He is currently working as an AI Scientist and Engineer at MindsLab. His research interests include financial machine learning, deep learning for tabular data, convex optimization, reinforcement learning, pattern matching and representation learning.

BYUNGCHUL KIM received his B.S degree in Physics from University of California Los Angeles, Los Angeles, CA, United States in 2016. He is currently working as a Chief Technology Officer at HighDev and as a Chief Data Officer at Synotex. His work involves architectural design of the research and development pipeline and his main research interests include financial machine learning, reinforcement learning, and practical application of machine learning in engineering perspective.

HEUISEOK LIM received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Korea University. His research interests include natural language processing, machine learning, and artificial intelligence.