Market Sentiment Analysis Based on Social Media and Trading Volume for Asset Price Movement Prediction

Jiahao Li, Yuyun Gong, Qinghua Zhao, Yufan Xie, Simon Fong, and Jerome Yen(✉)

Faculty of Science and Technology, University of Macau, Macau SAR, China
{mc05504,mc15076,mc05505,mc05384,ccfong,jeromeyen}@um.edu.mo

Abstract. As more and more netizens participate in financial market transactions, online discussions on asset price movements are becoming more comprehensive and timely. Online text, especially from social media, has the potential to be an important data source for financial opinion mining. Market sentiment analysis mainly includes direct analysis methods in the form of text-based surveys and indirect inference methods based on structured data such as price, trading volume, and volatility. In theory, the former is helpful for us to understand investor sentiment earlier, but due to the difficulty of obtaining a sufficient number of objective survey samples, its obtained research attentions are far less than the latter. To combine the advantages and offset the weakness of these two approaches, this paper uses Valence Aware Dictionary and Sentiment Reasoner (VADER) and Fast Fourier Transform (FFT) to construct social media sentiment indexes based on plenty of daily discussion texts about Bitcoin (BTC) and S&P500 (SPX) from Reddit for analyzing their interaction with prices. We also propose a new time series synchronization verification method called Rolling Time-lagged Cross-correlation (RTLCC) surface, and corresponding feature constructing methods, in which RTLCC helps us observe Time-lagged Cross-correlation from the perspective of Rolling Correlation while determining the hyperparameters (Window Size & Time Offset) for features construction. Finally, based on these features, we use four machine learning classifiers for modeling and verify the effectiveness of the proposed market sentiment analysis pipeline, in which on the prediction of 10-day price movements, the best model achieves 89.9% in accuracy (ACC) and 92.5% in AUC.

Keywords: Price Movement Prediction · Market Sentiment Analysis · VADER · Fast Fourier Transform · Time Series Synchronization Verification · Machine Learning

1 Introduction

Behavioral Finance Theory [17] argues that asset price movements align with the prevailing investors' sentiment. Since the text data of social media contains

X. Yang et al. (Eds.): ADMA 2023, LNAI 14176, pp. 383–398, 2023.
https://doi.org/10.1007/978-3-031-46661-8_26

abundant information that can be used to support trading decision-making [21], Market Sentiment Analysis is becoming a research hot spot in predicting financial market movements. The sentiment from social media text directly affects investors' propensity to trade, so it is often considered a direct measurement of market sentiment. Besides, the trading volume is proven to correlate with the market sentiment [15], and it is used to be an indirect measurement of market sentiment because it indirectly reflects the market sentiment. If the above two can be combined, not only the expansion of data sources is realized, but also the convenience from structured data analysis ideas of the latter are introduced while maintaining the timeliness of the former. Thus we attempt to quantify the correlation between market sentiment, trading volume, and price, aiming to construct a general pipeline for predicting price movement based on market sentiment analysis.

To be specific, we summarize market sentiment analysis approaches and introduce related works in Sect. 2. And in third section, a VADER-based sentiment index constructed on text data from Raddit discussion is proposed. Then we design a new synchronization verification indicator, RTLCC surface, and a set of relevant feature construction methods in the feature engineering. For the experimental part in Sect. 4, we respectively evaluate the effectiveness of these features on Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB), and LightGBM (LGB), while comparing their predicting performances. In short, the overall process has been briefly summarized in Fig. 1.

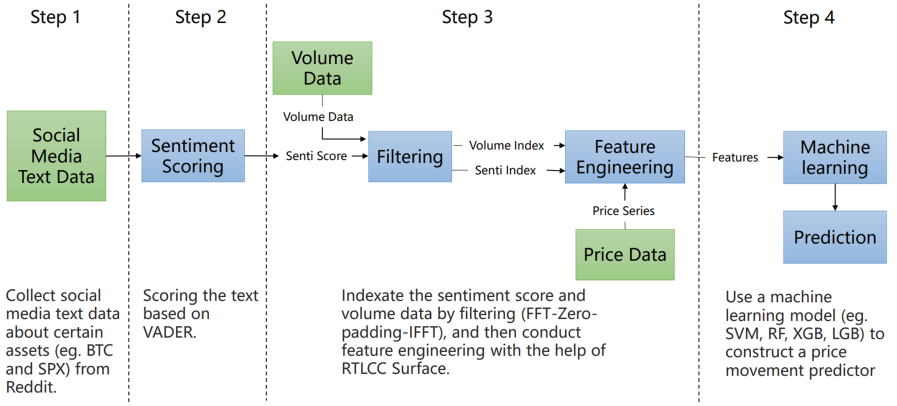
Step 1
Step 2
Step 3
Step 4
Social Media Text Data
Sentiment Scoring
Volume Data
Volume Data
Senti Score
Filtering
Volume Index
Senti Index
Feature Engineering
Price Series
Price Data
Features
Machine learning
Prediction
Collect social media text data about certain assets (eg. BTC and SPX) from Reddit.
Scoring the text based on VADER.
Indexate the sentiment score and volume data by filtering (FFT-Zero-padding-IFFT), and then conduct feature engineering with the help of RTLCC Surface.
Use a machine learning model (eg. SVM, RF, XGB, LGB) to construct a price movement predictor

Fig. 1. The overall pipeline of the asset price movement prediction

2 Related Work

2.1 Market Sentiment Analysis

Market Sentiment Analysis applies the NLP technology to quantitative finance and targets to analyze people's attitudes toward assets through computation of subjectivity in texts [22]. The analysis results are often turned into the sentiment index [19], which is a productive tool to quantify sentiment in figures and can be widely applied for Financial Market Predictive Analysis.

For the industry application of sentiment index, back in 1993, the CBOE Volatility Index (VIX) [20] was introduced to measure the market's expectation of 30-day volatility based on the assumption that the trading volume was a good proxy for investors' sentiment [6], and in nowadays, the S&P 500 Twitter Sentiment Index and S&P 500 Twitter Sentiment Select Equal Weight Index are always used to track the performance of the constituents with the most positive sentiment. Furthermore, a growing body of research keeps showing the actual value of the sentiment index: Huang et al. [9] devised an index capable of revealing investors' sentiment and predicting the overall stock market by using the least squares method, which outperformed well-established macroeconomic variables; Da Z Engelberg et al. [5] constructed a sentiment index to predict short-term returns and volatility, which is derived from daily Internet search volumes based on millions of households.

In addition, many studies also show that it is feasible and efficient to utilize the sentiment from social media to enhance financial data mining. Karabulut [11] declared that Facebook's Gross National Happiness (GNH) with the ability to predict changes in both daily returns and trading volume in the US stock market. As one of the most representative methods for sentiment analysis of social media, VADER [10] is an efficient rule-based algorithm that can help calculate a specified set of predetermined sentiment scores by identifying each feature (word, expression, and abbreviation) in a sentence. Toni Pano et al. [14] performed VADER-based sentiment analysis on BTC tweets to identify the role of different text preprocessing strategies in predicting Bitcoin prices. Kim Y B et al. [12] successfully predicted the price fluctuations of cryptocurrencies such as Bitcoin and Ethereum by using VADER to tag user comments in online communities.

2.2 Price Movement Prediction Based on Machine Learning

It is difficult to achieve financial prediction using market sentiment analysis alone, which is often only used as an important factor mining method. To achieve price movement prediction, it is also necessary to build a prediction model based on classification algorithms, and machine learning is a promising method. Nowadays, many scholars convince that some patterns are invisible to traditional financial or economic theories but can be detected and exploited by machine learning. Therefore, they have tried to use different machine-learning models to predict asset price movement.

As early as 2013, Alexander Porshnev et al. [16] used SVM with historical close price and Twitter tweets as input to achieve better results than random prediction in the price movements prediction of the S&P 500 Index and found that the market sentiment derived from text data improves the performance of the predictor. Furthermore, Al Nasseri et al. [1] demonstrate that decision tree algorithms can effectively quantify the relationship between semantic terms on StockTwits and trading behavior, like forecasting the impact of sentiment changes on the Dow Jones Industrial Average (DJIA) index, which has helped us to understand how emotions and language on social media platforms influence financial markets and provides a potential avenue for developing decision-making tools in the investment field. Recently, Guliyev H [8] compared four different machine learning models on predicting the monthly movements of WTI Oil's price, which shows that the XGBoost model made the best result of 91.8% accuracy.

3 Methodology

In this paper, VADER was used for calculating the sentiment score based on the discussions about different assets (e.g., BTC and SPX) on social media (Reddit) every day, which are Fourier transformed into sentiment indexes. Besides, Fourier Transform will also be performed on these assets' volumes to obtain another kind of sentiment index. Then, multiple time series correlation analyses, feature construction, and feature selection are performed on the price time series and the above two sets of sentiment indexes to obtain important features.

Finally, four different machine learning models (e.g., SVM, RF, XGB, LGB) are used to predict the movement of assets' prices with these features, and the prediction is a binary time series (e.g., up and down).

3.1 Sentiment Index Construction

As the information update frequency of the sentiment index is too fast to be matched with the price, it is a suitable way to change sentiment analysis into a sequence prediction task by constructing sentiment indexes. The data of close price and volume can be collected from Yahoo Finance as the price series and volume sequence. The sentiment analysis begins with the collection of daily discussion texts for specific assets from Reddit. The volume sequence can be used to represent the indirect sentiment index (Volume Index), while the direct sentiment index (Senti Index) needs to be constructed on these texts. The construction steps are as follows:

(1) Use API. PRAW to scrape the daily discussion describing a specific asset from Reddit. The statical description of the collected data is shown in Table 1.

Table 1. Data description.

	BTC	SPX
Time Frame	2018.3.1 - 2022.3.1	2018.3.1 - 2022.3.1
Trading Days	1455	1008
Comment Numbers	371477 (77752.881kb)	735289 (156367.301kb)
Missing Values/Days	86 (5.9%)	10 (1.0%)

(2) Calculate the sentiment scores of those text data by VADER and aggregate them by days. An example of how to get a sentiment score of a comment by VADER is shown in Table 2, in which the α is the approximate max sentiment score, and the meanings of the 'pos', 'neg', 'neu', 'total' are respectively the positive, negative, neutral, and compound sentiment score. +1 is to compensate for neutral words.

(3) Interpolate missing values by moving averages. Discussions on social media for certain assets are not present on all trading days, resulting in missing values in the sentiment score sequence. The numbers of various assets' missing values are shown in Table 1, none of which exceeds 5.9% of the respective trading days.

(4) Remove the noise of the sentiment scores by transforming this time-domain sequence into the frequency-domain one by FFT [4] and filtering out relative high-frequency components according to the threshold T while conducting zero-padding, which refers to the ratio of the relative low-frequency part retained after low-pass filter processing to all original components. Then the zero-padded sequences will be turned back into the time domain form by Inverse Fast Fourier Transformation (IFFT) [4]. For ease of representation, FFT_T represents a complete FFT-Zero-padding-IFFT period with $T\%$ as the filtering threshold like $2\%,\ 4\%,\ \ldots,\ 100\%$. Figure 2 describes the construction process of BTC's Senti Index when T is equal to 10%.

Table 2. An Example of VADER-based Sentiment Scoring.

	pos	neg	neu	total
Just			0+1	**Normalizing function:** $\frac{x}{\sqrt{x^2+\alpha}}$ $\mathbf{x = -2.4 - 2.1 = -4.5}$
gonna		0+1		
hurt		2.4+1		
the			0+1	
poor		2.1+1		
even			0+1	
more			0+1	
			0+1	
		6.5	6	12.5
	0/12.5	6.5/12.5	6/12.5	$-4.5/((-4.5^2)+15)^{0.5}$
	0	0.52	0.48	-0.7579

From top to bottom of Fig. 2, the first and last subplots respectively represent the time-domain sentiment sequence before and after processing. The second and third subplots represent the frequency-domain sentiment sequence before

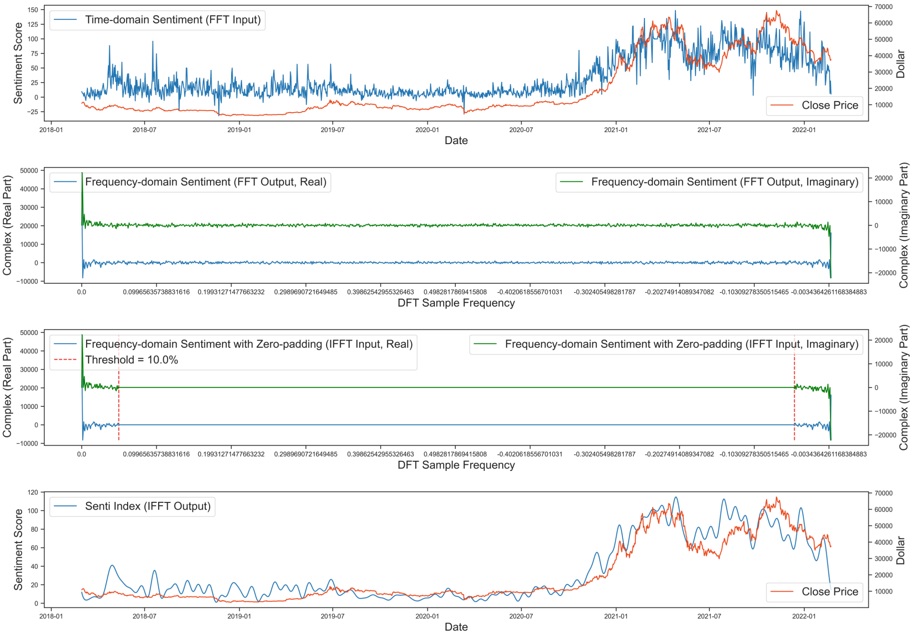

Fig. 2. The FFT-Zero-padding-IFFT period of BTC's Senti Index under FFT_{10}.

and after zero-padding respectively, and their x-axis is an array containing the Discrete Fourier Transform (DFT) sample frequency bin centers in cycles/second of the sample spacing, while their left and right y-axes represent the real and imaginary parts of the complex form of the frequency components.

3.2 Synchronization Verification

For ease of understanding, we declare definitions of universal variables for Sect. 3.2 and Sect. 3.3 uniformly on Table 3.

Pearson Correlation is a global measurement of the time series synchronicity, which calculates correlation by taking a linear relationship as one value between −1 and 1. It is easy to get an intuitive interpretation from Table 4 that there is a high correlation between the price and sentiment indexes. By introducing the WS, Rolling Pearson Correlation can calculate the Pearson Correlation in each rolling window, thus its measurement of the correlation is more comprehensive, but the leading relationships between sequences still cannot be observed. Based on Pearson Correlation, Time-lagged Cross-correlation (TLCC) [18] is used to determine which sequence is the leading sequence by introducing the TO. However, the above methods can not verify the synchronicity in a fine-grained way by observing the relationship among WS, TO, and $\widehat{C}$ at the same time. Windowed Time-lagged Cross-correlation (WTLCC) [3] combines WS and TO to calculate the TLCC in each fixed-size window, but its calculation result will be distorted due to the lack of data for actual calculations when the absolute value of TO is close to the size of the preset window.

This paper proposed a new synchronization verification method in time series analysis for the sentiment indexes designed above and related price time series, called the RTLCC surface, which can determine the values of WS and TO as hyperparameters for the subsequent feature construction while observing TLCC calculated in Rolling Correlation. The RTLCC is designed to calculate and find out the extreme value of the correlation by enumerating (WS, TO) combinations.

The enumerated values of TO construct the x-axis and the ones of WS are for the y-axis, whereas the z-axis is composited by $\widehat{C}$. Then a 3D coordinate map of $(TO,\ WS,\ \widehat{C})$, the RTLCC surface, is built up as Fig. 3.

Table 3. Variable definition declarations.

Name	Description
r_{sp}	The Pearson Correlation Coefficient between Senti Index and the price
r_{vp}	The Pearson Correlation Coefficient between Volume Index and the price
WS	The window size
TO	The time offset
C_{sp}	The Rolling Correlations between Senti Index and the price: $C_{sp} = \{c_{win}, c_{win+1},\ \ldots, c_i\},\ i \le n$ where n (same below) is the number of trading days and win (same below) is the size of the rolling window.
C_{vp}	The Rolling Correlations between Volume Index and the price
$\widehat{C}$	The average of Rolling Correlations between Senti/Volume Index and the price
c_i	The i^{th} value of C_{sp} or C_{vp}: $c_i = \frac{\sum_{j=i-win}^{i}(x_j-\hat{x})(y_j-\hat{y})}{\sqrt{\sum_{j=i-win}^{i}(x_j-\hat{x})^2}\sqrt{\sum_{j=i-win}^{i}(y_j-\hat{y})^2}}$ where x_j, y_j are the j^{th} sample pair in the sequences, while $\hat{x}$, $\hat{y}$ are relevant means.
MAC_{sp}	The moving average of C_{sp}: $MAC_{sp} = \{avg_{2\times win}, avg_{2\times win+1},\ \ldots, avg_i\}$ where i is less than or equal to the length of C_{sp}, while $avg_i = \frac{1}{win}\sum_{j=i-win}^{i} c_j$.
MAC_{vp}	The moving average of C_{vp}
MSC_{sp}	The moving standard deviation of C_{sp}: $MSC_{sp} = \{std_{2\times win}, std_{2\times win+1},\ \ldots, std_i\}$ where i is less than or equal to the length of C_{sp}, while $std_i = \sqrt{\frac{\sum_{j=i-win}^{i}(c_j-avg_i)}{win-1}}$.
MAC_{vp}	The moving standard deviation of C_{vp}

Table 4. Pearson correlation snapshot.

r_{sp}, r_{vp}	BTC	SPX
FFT_{10}	0.935, 0.815	0.928, −0.175
FFT_{20}	0.924, 0.794	0.904, −0.170
FFT_{30}	0.915, 0.766	0.892, −0.167
FFT_{40}	0.907, 0.754	0.881, −0.164
FFT_{50}	0.902, 0.743	0.873, −0.162
FFT_{60}	0.896, 0.729	0.862, −0.160
FFT_{70}	0.891, 0.721	0.852, −0.159
FFT_{80}	0.887, 0.716	0.844, −0.156
FFT_{90}	0.883, 0.711	0.838, −0.155
FFT_{100}	0.879, 0.703	0.829, −0.154

The construction processes of the surface are as follows:

(1) Select WS_j from $Range_{WS}$ like $3,\ 4,\ \ldots,\ 63$ as the rolling window size.
(2) Select TO_j from $Range_{TO}$ like $-30,\ -29,\ \ldots,\ 30$ as the offset based on the selected WS_i.
(3) Loop through the above two steps in a nested way and compute all

$$\widehat{C}_{ij} = \frac{1}{n - WS_i} \sum_{k=WS_i}^{n-j} c_{kj}. \tag{1}$$

of which every value c_i of corresponding Rolling Correlations is adjusted to be

$$c_{kj} = \frac{\sum_{p=k-WS_i}^{k} \left(x_p - \hat{x}_k\right)\left(y_{p+j} - \hat{y}_k\right)}{\sqrt{\sum_{p=k-WS_i}^{k} \left(x_p - \hat{x}_k\right)^2}\sqrt{\sum_{p=k-WS_i}^{k} \left(y_{p+j} - \hat{y}_k\right)^2}}. \tag{2}$$

in which x_p is p^{th} value in Senti Index or Volume Index, while y_{p+j} is $(p+j)^{th}$ value in the relevant price sequence. And $\widehat{x}_k = \frac{1}{WS_i} \sum_{p=k-WS_i}^{k} x_p$, while $\widehat{y}_k = \frac{1}{WS_i} \sum_{p=k-WS_i}^{k} y_{p+j}$ when the j is fixed.

Each point on the RTLCC surface represents the average of a certain group of Rolling Correlations between the index and price under a specific combination of WS_i and TO_i, and this kind of average is also adopted in studies [2,7] as an important financial indicator. An ideal coordinate point can be defined in this 3D space when the meaning of each axis has been defined in directions. For example, the ideal point can be $(\ WS_{id},\ TO_{id},\ \widehat{C}_{id}) \rightarrow (0,\ -\infty,\ +\infty)$ to make the 'correlation' $\widehat{C}_{ij}$ positively highest possible with the smallest window size and most negative offset (Scene 1). The reasons for pursuing a small window and negative offset are: a smaller window means spending less time looking back at the historical data, so there will be fewer data that has to be used for the model learning; a more negative offset means that the more days the sentiment index leads the price, so there will be longer periods of time can be used to design trading strategies in advance. Moreover, the ideal point also can be ($WS_{id},\ TO_{id},\ \widehat{C}_{id}) \rightarrow (0,\ -\infty,\ -\infty)$ to make the 'correlation' $\widehat{C}_{ij}$ negatively highest possible with the smallest window size and most negative offset (Scene 2), when there is a potential negative correlation between the sequences to be observed.

This paper adapts the Weighted Euclidean Distance based on Min-Max Normalization to calculate the distance from the ideal point to every point on the surface, which enables the comparability between these three dimensions and the one between surfaces. And in Scene 1, the ideal point is $(WS_{id}, TO_{id}, \widehat{C}_{id}) = (0, 0, 1)$, while the ideal point becomes $(0, 0, 0)$ in Scene 2. Denote $(WS_i,\ TO_j, \widehat{C}_{ij})$ as 3D coordinates of any point on the surface, its distance to the idea point is:

$$Distance\left(WS_i,\ TO_j, \widehat{C}_{ij}\right) = \sqrt{w_1 \times (WS_i - WS_{id})^2 + w_2 \times \left(TO_j - TO_{id}\right)^2 + w_3 \times \left(\ \widehat{C}_{ij} - \widehat{C}_{id}\right)^2}. \quad (3)$$

in which the w_1, w_2, and w_3 are weights representing the importance of each parameter, subjecting to $w_1 + w_2 + w_3 = 1$. Ultimately, the entire analysis task is reduced to finding the minimum of the 2D matrix. For instance, the RTLCC of Bitcoin under FFT_{10} is shown in Fig. 3, from which three hyperparameters of the actual best point are obtained and recorded in Table 5.

Table 5. Best parameter combination for BTC under FFT_{10} found by RTLCC based on the training set.

	Volume Index VS Bitcoin Price	Senti Index VS Bitcoin Price
Window Size ($WS^{'}$)	55 days	56 days
Time Offset ($TO^{'}$)	2 days	6 days
Rolling Correlations' Average ($\widehat{C}^{'}$)	0.331	0.522

3.3 Feature Construction

In this subsection, we construct features based on the price, Senti Index, and Volume Index. Then hundreds of features for each asset are constructed, which can be summarized into 28 features according to different FFT_T, as shown in Table 6, where price_up_down is the predicting target.

This paper regards price movement prediction as a multivariate time series forecasting task and believes each asset price depends not only on its historical values but also on its relationship with relevant sequences. Thus, features are constructed based on two rules: the construction based on short-term and long-term change value in single time series (Rule 1); the construction based on correlations between various time series (Rule 2).

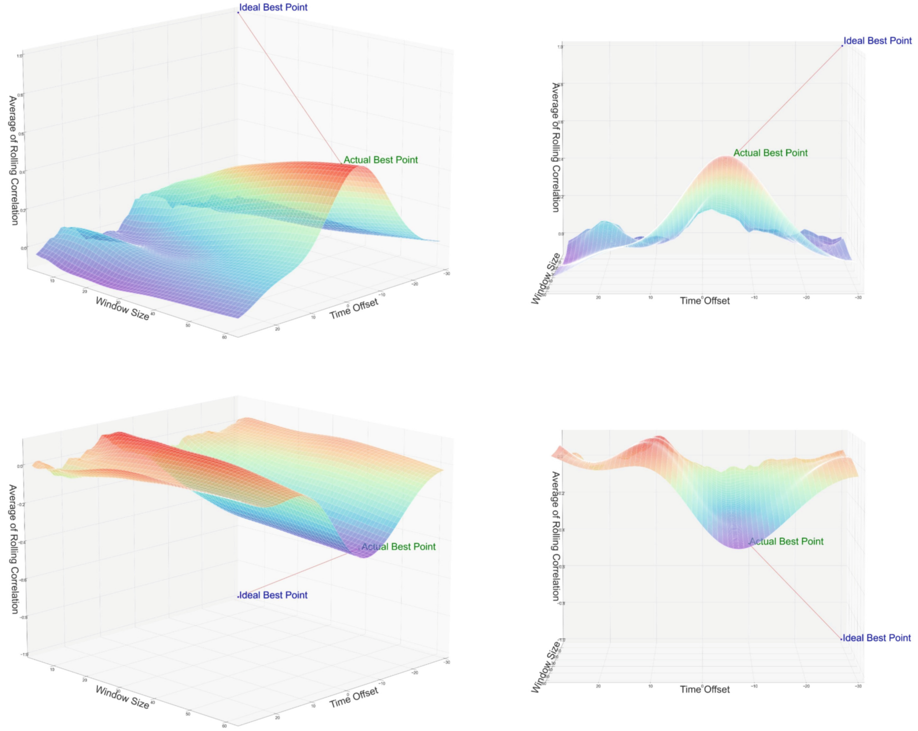

Fig. 3. The RTLCC surfaces of BTC's Senti Index (upper two) and SPX's Volume Index (bottom two) under FFT_{10} based on the training set.

Based on Rule 1, features 1 to 15 are constructed. The short-term trend refers to the difference between two consecutive sequence units, while the long-term trend refers to the trend in a specific time span having more than two units. For the price $P_A = \{p_1,\ p_2,\ \ldots,\ p_i\},\ \ i \leq n$, the first order difference, price_diff, is equal to $p_i -\ p_{i-1}$, and the short-term trend, price_trend, is computed by setting $x = p_i -\ \ p_{i-1}$ in this formula:

$$\mathrm{Sign}(x) = \begin{cases} 0,\, x \leq 0 \\ 1,\, x > 0 \end{cases}. \tag{4}$$

The long-term change value, his_price_up_down_value, is represented by the tangent of the included angle α between the first-order polynomial fitted straight line $y_1 = A \bullet x + B$ with the horizontal line $y_2 = B$. When the x is equal to TO', this value is:

$$tan\alpha = \frac{y_1|_{x=TO'} - y_2}{x} = \frac{A \bullet TO' + B - B}{TO'} = A. \tag{5}$$

Meanwhile, the long-term trend, his_price_up_down_trend, is defined as $Sign(\tan\alpha - tan0°)$. For both Senti Index and Volume Index, the definition of

Table 6. Feature description.

	FEATURES	DESCRIPTION
0	price_up_down	Future 1, 5, or 10 days' price movement (0, 1)
1	price	Price time series
2	price_diff	The amount of price change between two consecutive days
3	price_trend	Price's short-term trend $Sign(price\_diff)$ (0, 1)
4	his_price_up_down_value	The amount of price change in the past TO' days $tan\alpha$
5	his_price_up_down_trend	Price's long-term trend in the past TO' $Sign(tan\alpha)$ days (0, 1)
6	senti (FFT_T)	Senti Index under FFT_T
7	senti_diff (FFT_T)	The amount of change in the Senti Index for two consecutive days
8	senti_trend (FFT_T)	Senti Index's short-term trend $Sign(senti\_diff)$ (0, 1)
9	his_senti_up_down_value (FFT_T)	The amount of Senti Index's change in the past TO' days
10	his_senti_up_down_trend (FFT_T)	The trend of Senti Index's change in the past TO' days (0, 1)
11	volume (FFT_T)	Volume Index under FFT_T
12	volume_diff (FFT_T)	The amount of change in Volume Index between two consecutive days
13	volume_trend (FFT_T)	Volume Index's short-term trend $Sign(senti\_diff)$ (0, 1)
14	his_volume_up_down_value (FFT_T)	The amount of Volume Index's change in the past TO' days
15	his_volume_up_down_trend (FFT_T)	The trend of Volume Index's change in the past TO' days (0, 1)
16	rolling_corr_senti (FFT_T)	Senti Index's rolling correlations C_{sp}
17	mean_std_diff_senti (FFT_T)	The difference between MAC_{sp}, the moving average of C_{sp}, and MSC_{sp}, the moving standard deviation of C_{sp}
18	mean_std_sum_senti (FFT_T)	The sum of MAC_{sp} and MSC_{sp}
19	corr_up_value_senti (FFT_T)	The difference between C_{sp} and (MAC_{sp}+MSC_{sp})
20	corr_down_value_senti (FFT_T)	The difference between C_{sp} and (MAC_{sp}-MSC_{sp})
21	corr_up_down_senti (FFT_T)	The state of C_{sp} (−1, 0, 1)
22	rolling_corr_volume (FFT_T)	Volume Index's rolling correlations C_{vp}
23	mean_std_diff_volume (FFT_T)	The difference between MAC_{sp}, the moving average of C_{vp}, and MSC_{sp}, the moving standard deviation of C_{vp}
24	mean_std_sum_volume (FFT_T)	The sum of MAC_{vp} and MSC_{vp}
25	corr_up_value_volume (FFT_T)	The difference between C_{vp} and (MAC_{vp}+MSC_{vp})
26	corr_down_value_volume (FFT_T)	The difference between C_{vp} and (MAC_{vp}-MSC_{vp})
27	corr_up_down_volume (FFT_T)	The state of C_{vp} (−1, 0, 1)

short-term or long-term features are similar. For the predicting target, the asset price movements are defined as price_up_down, of which the formula is:

$$Movement = Sign\left(tan\alpha\right) = Sign\left(\frac{y_1|_{x=future} - y_2}{x}\right). \tag{6}$$

where $future$ can be any reasonable integer such as 1, 5, 10, which means the time lag of each predicting step will be 1, 5, or 10 trading days.

Based on Rule 2, features 16 to 27 are constructed. The Bollinger Bands strategy is adopted to observe the correlation between the sentiment index and the price of the same asset based on the assumption that sentiment can reflect future prices in advance. There are five sequences concerned in our Bollinger Bands strategy, namely the normalized price sequence, normalized Senti/Volume index, Rolling Correlations between price and the normalized index, and the upper and lower boundary of the Bollinger Band. By observation, we find that future price movements are often related to the current trends of both the price and sentiment index, as well as the relationship between the rolling correlations and the upper and lower boundary. To quantify the association between these sequences, the following feature construction is performed.

Take Senti Index as an example, the *win* of variables in Table 3 equals to $WS^{'}$ and the state of C_{sp}, corr_up_down_senti, can be formulated as:

$$\mathrm{State}\left(C_{sp}\right) = \begin{cases} 1, & \mathrm{state1} = 1 \\ 0, & \mathrm{state1} = 0, \mathrm{state2} = 1 \\ -1, & \mathrm{state1} = 0, a\mathrm{state2} = 0 \end{cases}. \tag{7}$$

$$state1 = Sign\left[C_{sp} - \left(MAC_{sp} + MSC_{sp}\right)\right]. \tag{8}$$

$$state2 = Sign\left[C_{sp} - \left(MAC_{sp} - MSC_{sp}\right)\right]. \tag{9}$$

As a result, the above formula will give out a ternary sequence, and the relationship between the sentiment index and price is summarized into three types of situations. Finally, features [1, 2, 4, 6, 7, 9, 11, 12, 14, 16–20, 22–26] need to be standardized into the zero-mean and unit variance distribution as:

$$v' = \frac{v - \mathrm{mean}(V)}{\sigma}. \tag{10}$$

where v is every single value in a certain feature sequence V, while σ is the standard deviation of V.

4 Evaluation

The evaluation process is carried out according to an experiment comparing the classification performance of four different machine learning models constructed based on two different sets of features as the training data.

4.1 Validation Method and Indicator

Walk-forward Validation method is used in this paper, which adopts the sliding method to split the training set and the test set, and it only takes the part of the accessible historical data closest to the predicted time span as the training set. Moreover, the indicators used here are the 0.5-threshold Accuracy (ACC) and Area under the ROC Curve (AUC).

In the prediction process, the errors will continue to accumulate. The larger the test set is divided, the harder for steps at the end of the test set to be predicted accurately. Thus, the sub-test set with a small size of just one predicting step for each test iteration was designed in the following experiment, and all predicted values are then concatenated in chronological order for comparison with the target values. The overall test set (out-of-sample data) consisting of every sub-test set accounts for 20% of all data and the training set (in-sample data) used for each prediction always accounts for 80% of all data.

4.2 Experiment

In the comparative experiment, four different machine learning models, including SVM, Random Forest, XGB, and LGB are used to be classifiers with the Original or All Features as the training data. Then, by comparing the ACC and AUC of these classifiers, it is proved that the proposed feature construction method is effective, and the difference in the performance of models is further explored with results in Table 7. The 'Indicator' and 'Change' are expressed as percentages, in which the former represents the value of ACC or AUC under a specific combination of training data, model type, and time lag per predicting step, while the latter represents the change of the above indicator values upon their own mean. The 'Mean' is the mean of the four models' performance under the same time lag and training data. In addition, the Average Change represents the average of the Changes under different training data and the same time lag and model type. The Original Features and All Features, respectively, refer to modeling by using only five original features (features 1 to 5) and using all features (features 1 to 27).

The Original (ori_) or All (all_) in Fig. 4 represents that the Original or All Features are training data, and the polyline with a name containing the above two abbreviations is associated with the performance of the corresponding model. Among them, the red line and the blue line represent the moving average of ACC with 10% of the test set's size as the rolling window size, and they are used to observe the change in the predicting accuracy of the model over time. The four parallel lines represent the global metrics of the entire test set, and their specific scores are shown in the title of the figure.

Through the above experiments, under the same training data, whether for ACC or AUC, it can be found that the model based on the Original Feature is inferior to the corresponding model based on All Features, and they all increase with a longer time lag. This shows that our feature construction method is able

Table 7. Comparative experiments results

Trainning Data		Original Features (BTC)		All Features (BTC)		Original Features (SPX)		All Features (SPX)		Average Change
% Indicator (Change)		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	
Time lag/Step	Asset									
1 day	SVM	48.7 (4.5)	47.1 (1.9)	57.4 (0.0)	58.5 (−3.9)	52.6 (6.7)	40.4 (−6.7)	55.0 (3.6)	42.7 (−17.2)	−1.4
	RF	47.5 (1.9)	48.1 (4.1)	54.4 (−5.2)	56.5 (−7.2)	47.4 (−3.9)	42.6 (−1.6)	57.3 (7.9)	56.4 (9.3)	0.7
	XGB	46.0 (−1.3)	46.6 (−3.5)	57.8 (0.7)	62.5 (2.6)	45.6 (−7.5)	44.1 (1.8)	50.3 (−5.3)	52.9 (2.5)	−1.3
	LGB	44.1 (−5.4)	45.1 (−2.4)	60.1 (4.7)	66.1 (8.5)	51.5 (4.5)	45.9 (6.0)	49.7 (−6.4)	54.4 (5.4)	1.9
Mean		46.6	46.2	57.4	60.9	49.3	43.3	53.1	49.7	-
5 days	SVM	49.6 (−13.7)	54.5 (−8.6)	70.2 (−7.6)	78.9 (−4.4)	58.8 (−1.5)	59.2 (−4.4)	68.8 (−10.8)	76.3 (−5.5)	−7.1
	RF	58.0 (0.9)	58.6 (−1.7)	79.8 (5.0)	83.1 (0.7)	58.8 (−1.5)	65.1 (5.2)	80.6 (4.5)	82.9 (2.7)	2.0
	XGB	63.0 (9.6)	63.3 (6.2)	78.6 (3.4)	84.0 (1.8)	59.4 (−0.5)	61.3 (−1.0)	78.8 (2.2)	82.5 (2.2)	3.0
	LGB	59.2 (3.0)	61.9 (3.9)	75.2 (−1.1)	84.1 (1.9)	61.8 (3.5)	61.9 (0.0)	80.0 (3.8)	81.1 (0.5)	1.9
Mean		57.5	59.6	76.0	82.5	59.7	61.9	77.1	80.7	-
10 days	SVM	53.6 (−3.4)	53.5 (−5.8)	78.5 (−10.0)	86.1 (−4.2)	59.2 (−0.2)	59.4 (−1.2)	81.7 (−6.1)	85.8 (−5.3)	−4.5
	RF	52.9 (−4.7)	51.1 (−3.0)	89.7 (2.9)	90.2 (0.3)	58.0 (−2.2)	62.2 (3.5)	86.4 (−0.7)	91.0 (0.4)	−0.4
	XGB	57.5 (3.6)	58.5 (3.0)	89.7 (4.1)	90.2 (2.7)	58.0 (−2.2)	59.9 (−0.3)	89.9 (3.3)	93.0 (2.6)	2.1
	LGB	57.9 (4.3)	60.0 (5.6)	89.7 (2.9)	91.0 (1.2)	62.1 (4.7)	58.7 (−2.3)	89.9 (3.3)	92.5 (2.1)	2.7
Mean		55.5	56.8	87.2	89.9	59.3	60.1	87.0	90.6	-

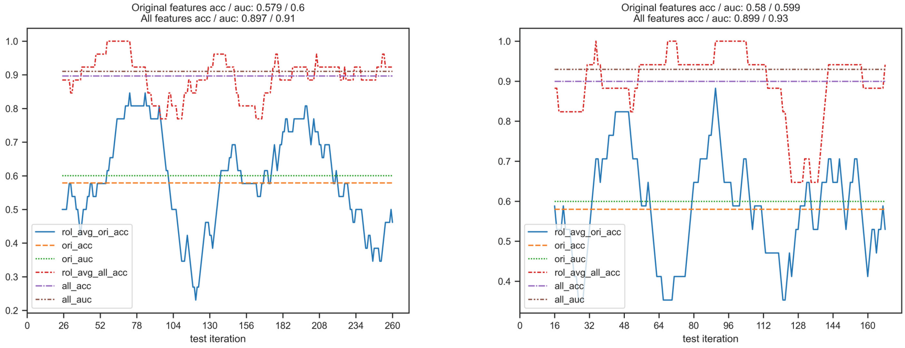

Fig. 4. The model comparison between Original Features and All Features of BTC-LGB under 10-day time lags (left) and the one of SPX-XGB (right).

to increase the classification's performance. For both Bitcoin and S&P 500, LGB and XGB models perform better on comprehensive performance than SVM and RF, no matter what the predicting time lag is equal to. Among them, the XGB model performs best when the time lag is equal to 5 days, with an Average Change of 3.0%, while the LGB model performs best when the time lag equals 10 days, with an Average Change of 2.7%. Although when the time lag is equal to 1 day and the training data is All Features, RF performs better both on ACC and AUC than other models, its performances are lower under any other same conditions. Moreover, by comparing the Average Change, even though XGB achieves the highest average improvement of 3% when the time lag equals 5 days, it performs below the average of all models in a one-day scenario, while only LGB is able to achieve the highest or second high average improvement in all cases, thus LGB is stated to be the best and most robust model founded in this experiment.

When we compare the performance of the models trained on all Features, we can also find that the related models of both BTC and SPX perform well on a 10-day-long time lag (Mean: ACCs' is 87.1%, and AUCs' is about 90.3%). However, the former has a clear advantage over the latter in a 1-day-short time lag. When the latter is almost unpredictable (Mean: ACC's is 53.1%, and AUC's is 49.7%), the former's means of ACC and AUC are respectively as high as 57.4% and 60.9%. It can be seen that the market sentiment of different assets has different effecting time lag on their price movements. In the short term, BTC is more effective than SPX, and as the time lag increases, the gap between the two continues to decrease.

5 Conclusions and Future Work

A novel pipeline of asset price movement prediction based on market sentiment analysis is proposed in this paper, which effectively solves the problem that the cost of text data collection by the direct analysis method of market sentiment is too high. At first, we quantify the market sentiment from social media and trading volume as sentiment indexes by VADER and FFT-Zero-padding-IFFT processing, and then change it into a form of structured data analysis while introducing the convenience of indirect analysis methods. Then, we design a new synchronization verification method, RTLCC, to find out the best observing window size and time offset for feature construction. Finally, classifiers for asset price movement prediction are built and compared based on four different machine learning algorithms for providing an experimental basis for model selection. In short, our research proves that it is feasible to use the public sentiment from social media for price movement prediction, and at the same time provides new ideas for unstructured data analysis for market sentiment, as well as methodological innovations in this field.

In the future, in terms of price movement prediction modeling, we plan to design ensemble models for further comparative experiments and more effective forecasting. For applications, we will try to increase the update frequency of both price and text to 30 or even 15 min among more assets to further explore the potential of this kind of model in high-frequency trading.

Acknowledgment. This work is supported by the Macau SAR Science and Technology Development Fund (No. 0032/2022/A, No. 0091/2020/A2), Guangzhou Development Zone Science and Technology Grant (No. 2021GH10, No. 2020GH10, and No. EF003/FST-FSJ/2019/GSTIC), and Collaborative Research Grant of University of Macau (No. MYRG-GRG2022, No. MYRG-CRG2021-00002-ICI).

References

1. Al Nasseri, A., Tucker, A., De Cesare, S.: Quantifying stocktwits semantic terms' trading behavior in financial markets: an effective application of decision tree algorithms. Expert Syst. Appl. **42**(23), 9192–9210 (2015)
2. Aloui, C., Nguyen, D.K., Njeh, H.: Assessing the impacts of oil price fluctuations on stock returns in emerging markets. Econ. Model. **29**(6), 2686–2695 (2012)
3. Boker, S.M., Rotondo, J.L., Xu, M., King, K.: Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychol. Methods **7**(3), 338 (2002)
4. Cochran, W.T., et al.: What is the fast fourier transform? Proc. IEEE **55**(10), 1664–1674 (1967)
5. Da, Z., Engelberg, J., Gao, P.: The sum of all fears investor sentiment and asset prices. Rev. Financ. Stud. **28**(1), 1–32 (2015)
6. Gervais, S., Kaniel, R., Mingelgrin, D.H.: The high-volume return premium. J. Financ. **56**(3), 877–919 (2001)
7. Goetzmann, W.N., Li, L., Rouwenhorst, K.G.: Long-term global market correlations (2001)
8. Guliyev, H., Mustafayev, E.: Predicting the changes in the WTI crude oil price dynamics using machine learning models. Resour. Policy **77**, 102664 (2022)
9. Huang, D., Jiang, F., Tu, J., Zhou, G.: Investor sentiment aligned: a powerful predictor of stock returns. Rev. Financ. Stud. **28**(3), 791–837 (2015)

10. Hutto, C., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
11. Karabulut, Y.: Can facebook predict stock market activity? In: AFA 2013 San Diego Meetings Paper (2013)
12. Kim, Y.B., et al.: Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PLoS ONE **11**(8), e0161197 (2016)
13. Oliveira, N., Cortez, P., Areal, N.: The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst. Appl. **73**, 125–144 (2017)
14. Pano, T., Kashef, R.: A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the ERA of COVID-19. Big Data Cogn. Comput. **4**(4), 33 (2020)
15. Pettengill, G.N.: Holiday closings and security returns. J. Financ. Res. **12**(1), 57–67 (1989)
16. Porshnev, A., Redkin, I., Shevchenko, A.: Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In: 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 440–444. IEEE (2013)
17. Ritter, J.R.: Behavioral finance. Pacific-Basin Financ. J. **11**(4), 429–437 (2003)
18. Shen, C.: Analysis of detrended time-lagged cross-correlation between two nonstationary time series. Phys. Lett. A **379**(7), 680–687 (2015)
19. Solt, M.E., Statman, M.: How useful is the sentiment index? Financ. Anal. J. **44**(5), 45–55 (1988)
20. Whaley, R.E.: The investor fear gauge (2000)
21. Xing, F.Z., Cambria, E., Welsch, R.E.: Natural language based financial forecasting: a survey. Artif. Intell. Rev. **50**(1), 49–73 (2018)
22. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **8**(4), e1253 (2018)