1. Introduction

The stock market refers to the collection of markets and exchanges where regular activities of buying, selling and issuance of shares of publicly-held companies take place. Buying and selling of shares completely depends on the market movement. Forecasting stock market trends has been treated as one of the most challenging but important tasks because of its nonlinear and dynamic behavior. Investor sentiment plays an important role on the stock market. Investor sentiment constitutes a key factor of the financial market. Textual content on the internet provides a precious source to reflect investor sentiment and predicts stock prices as a complement to traditional stock market time series data. Hence an automated approach is required to distill knowledge from a large number of textual documents. Sentiment analysis is used to automatically extract views, attitudes, and emotions from the opinionated contents. So, we employ sentiment analysis to sentiment indexes, and then aggregate them with stock data forecast movement direction.

In order to get an efficient and persuasive sentiment index, we take the day-of-week effect into consideration, which means that the average return on Mondays is much lower than that on the other days of the week. It is one of the most well-known financial anomalies dating back to 1930 when Fred C. Kelly revealed the phenomenon on the U.S. markets where the returns had the tendency to decline on Mondays. Then, the effect is proved to exist in global stock markets.

Another difficulty in predicting stock movement direction is attributed to its nonlinear, dynamic, and evolutionary properties. Support vector machine (SVM) has been widely utilized since it can solve the nonlinear problem by converting it to a quadratic programming. Moreover, the solution of SVM is unique and globally optimal. It can also reduce the over-fitting problem by selecting the maximal margin hyperplane in the feature space. To further address the problem, we implement five-fold cross validation. However, it leads to look-ahead bias, so we integrate SVM with a realistic rolling window approach to eliminate the bias. Empirical results illustrate that combining sentiment features with stock market data outperforms using only stock market data in forecasting movement direction.

2. Proposed Methodology

The aims is to forecast stock market movement direction by not only using financial market data, but also combining them with sentiment features that incorporate investor psychology. The features are extracted from unstructured news data automatically and then are expressed as sentiment indexes. In order to make the indexes more realistic and reliable, we take the day-of-week effect into consideration. Next, we employ SVM to forecast stock market trends, and make an adjustment to real market situations by use of a rolling window approach, and then compare the accuracy with the baseline method. Moreover, the prediction results are used to instruct investment decisions, and the performance of three different trading strategies are evaluated and compared. The overview of the stock market prediction architecture is illustrated in Fig-1.

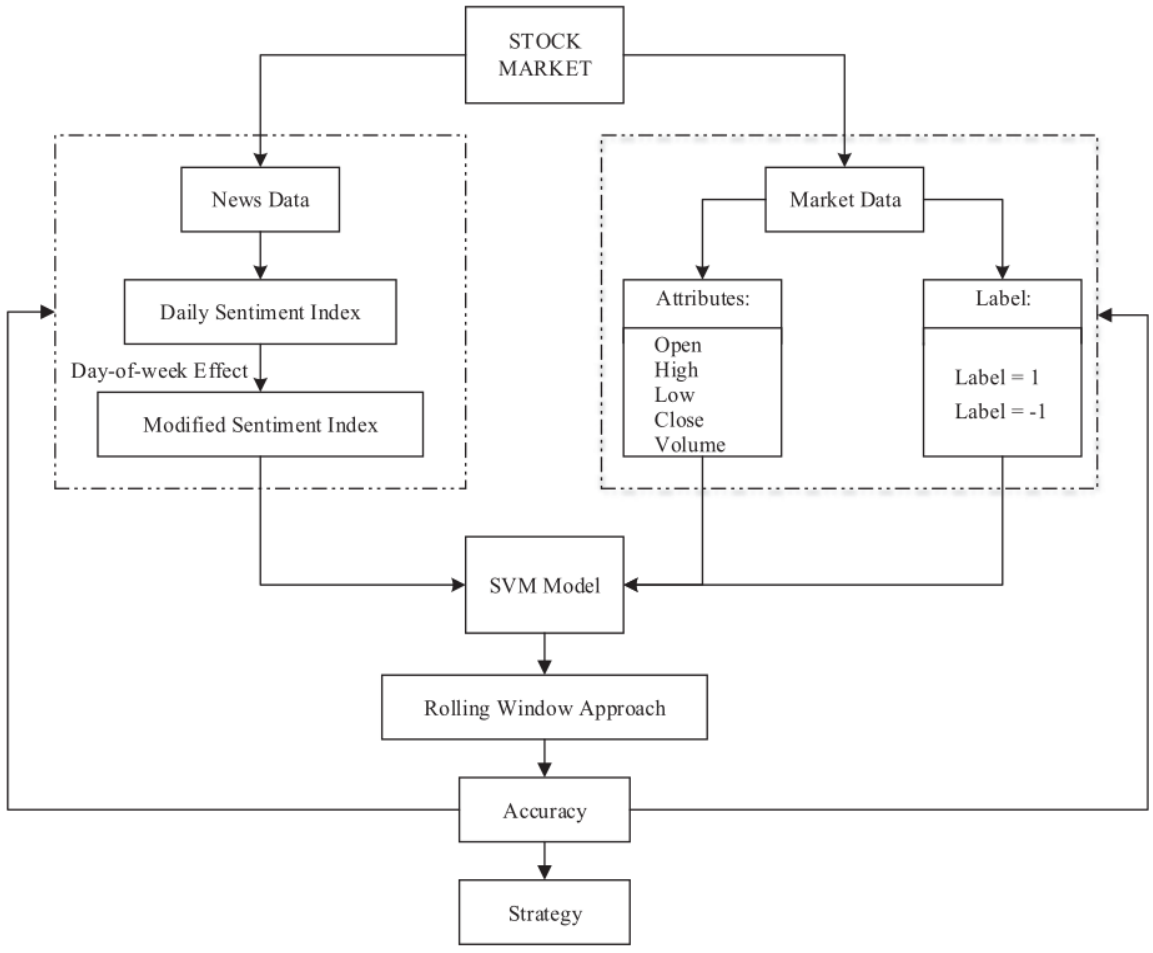

Fig-1: Overview of stock market prediction architecture.

The entire procedure of forecasting stock market is divided into two parts

A. Investor Sentiment

B. Support Vector Machine

A. Investor Sentiment

This section is made up of three steps. We first build a web crawler to download news documents automatically from the Internet, and then construct daily sentiment indexes based on the corpus. At last, adjustments are made in consideration of the day-of-week effect. Web crawler framework is explained in Fig-2.

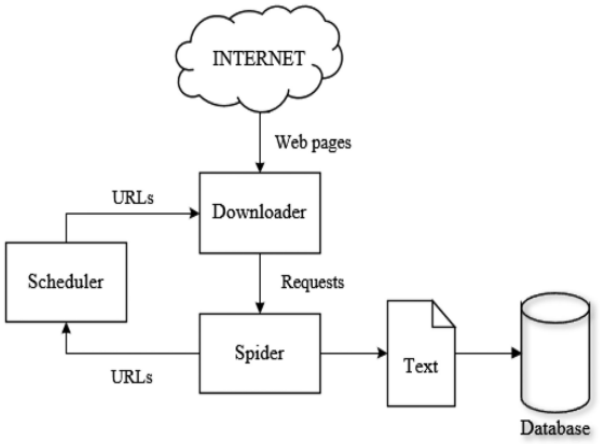

Fig-2: Framework of the web crawler.

Sentiment analysis will be performed on the downloaded text data to extract the investor's sentiment.

B. Support Vector Machine

SVM is a supervised machine learning model for classification, which was proposed by Vapnik in the 1990s. Assume that there is an input space X, an output space Y , and a training dataset T.

$$T = \{(x_i , y_i), i = 1, \ldots , l\} \in (X \times Y)^l$$

where $x_i \in R$, $y_i \in Y = \{-1, 1\}$. Here y=-1 means sell the shares and y=1 means buy shares.

3. Discussion and Result

We intend to explore the trend of a very important index in China, the SSE 50 Index, not only by using stock market data but also exploiting news documents related to it and its constituents. The SSE 50 Index is a primarily blue-chip stock index on the Shanghai stock market, and it is made up of the 50 largest stocks of good liquidity and representatives. Conventional time series data include opening price, closing price, high for the day, low for the day, trading volume in number of shares, trading volume in RMB, change in RMB, and change in percentage.

For sentiment analysis first segment each document into several sentences by identifying punctuation such as "." "," etc and then sentences are divided into separate words, and if there appears a negative word, it is treated as a whole with the word next to it. Because single word may note be able to give true meaning. Then, we need to categorize each document, assume there are pi positive sentences and n_i negative sentences in document i; if $p_i > n_i$, the document is positive; if $p_i = n_i$, the document is neutral; if $p_i < n_i$, the document is negative.

Next, we implement two experiments to predict the index movement direction. Experiment 1 is to use market data, which include opening price, closing price, high for the day, low for the day, trading volume in number of shares, trading volume in RMB, change in RMB, and change in percentage. And then, we combine them with sentiment features for Experiment 2. We employ classification accuracy Acc to assess the performance, as shown in the following equation:

Empirical results illustrate that the accuracy of forecasting the movement direction of the SSE 50 Index can be as high as **89.93%** with a rise of **18.6%** after introducing sentiment variables. And, meanwhile, our model helps investors make wiser decisions. These findings also imply that sentiment probably contains precious information about the asset fundamental values and can be regarded as one of the leading indicators of the stock market.

4. References

[1] I. Perikos and I. Hatzilygeroudis, “Recognizing emotions in text using ensemble of classifiers”, *Eng, Appl. Artif. Intell.,* Vol. 51, pp. 191-201, 2016

[2] J. Boolen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.,* vol.2, no. 1, pp. 1-8, 2011.

[3] J. Zhang, Y. Lai, and J. Lin, “The day-of-week effects of stock markets in different countries,” *Finance Res. Lett.,* vol. 20, pp 47-62, 2017.