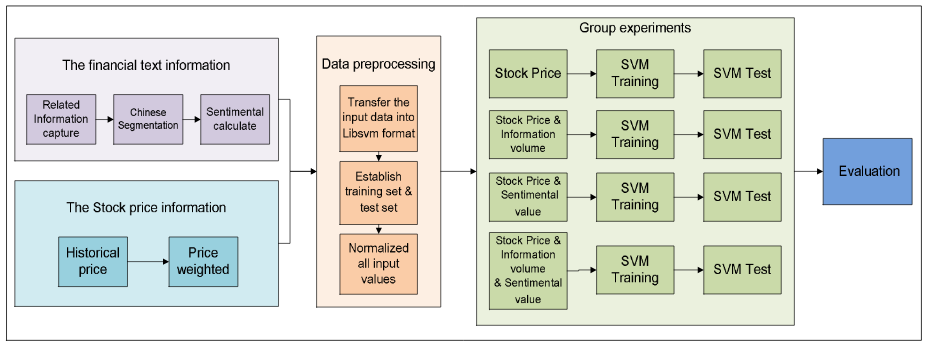
Stock Price Forecasting with Support Vector Machines Based on Web Financial Information Sentiment Analysis → Review

**Introduction**

The financial markets are a popular topic of study and price forecasting of stocks is one of the most difficult problems in time series prediction due to the noisy, unstructured and dynamic nature of the data. The internet makes it easier to get and discuss information about companies and markets faster and since information is a source of alpha or edge to some investors it makes outperforming the market more difficult. Most financial data is unstructured text and can be argued that the text has 2 main dimensions, volume and sentiment. This is why sentiment analysis is such an import topic of study, for financial markets sentiment analysis can provide an insight to the emotions and behaviors of market participants.

The study looks at three stocks in the Chinese stock market using sentiment analysis to create a sentiment value which is then combined with the price data. From there a model is built to forecast future prices using Support Vector Machines.

**Model and Method**

The model is comprised of 4 components show in the figure.

1. **The collection and computation** of financial text information and historical prices.

2. **Data pre-processign**, formatting the data, creating the training and test sets, and normalizing all inputs.

3. **Group experiments,** split the data into just price, price and information volume, price and sentiment value, and finally all 3, price, information volume, and sentiment value.

4. **Evaluating the results**

In the experiment 3 stocks, Toread, Hanwei Electronics, and Huayi Brothers, where randomly selected from the Growth Enterprise Market in China. Next a financial news crawler was used to gather data and the amount of news stories was used to create the volume feature.

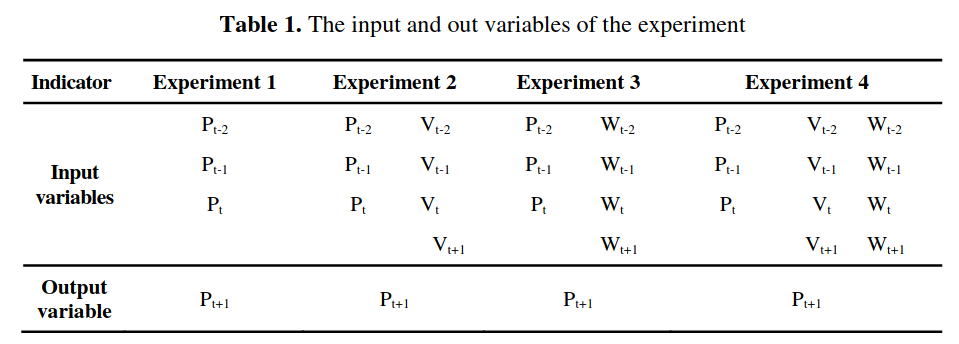
Calculating the sentiment value, for each news article the content and intensity was examined and the attitude, bullish, bearish and neutral. To find the intensity researchers looked at the effect the piece had on the market, ranging from high to low. The value is the overall score where positive scores indicate bullish and negative indicate bearish, and the absolute value of the score is the intensity. To create the score a bag of words approach was used where the words of the text are tockenized, stop words are removed and then the words are compared to a positive and negative word list.

From there historical price data was gathered from Yahoo Finance for the year 2010 and for Huayi Brothers there was an increase in shares so the price was adjusted. The price data was an average of daily open, close, high and low prices for that day.

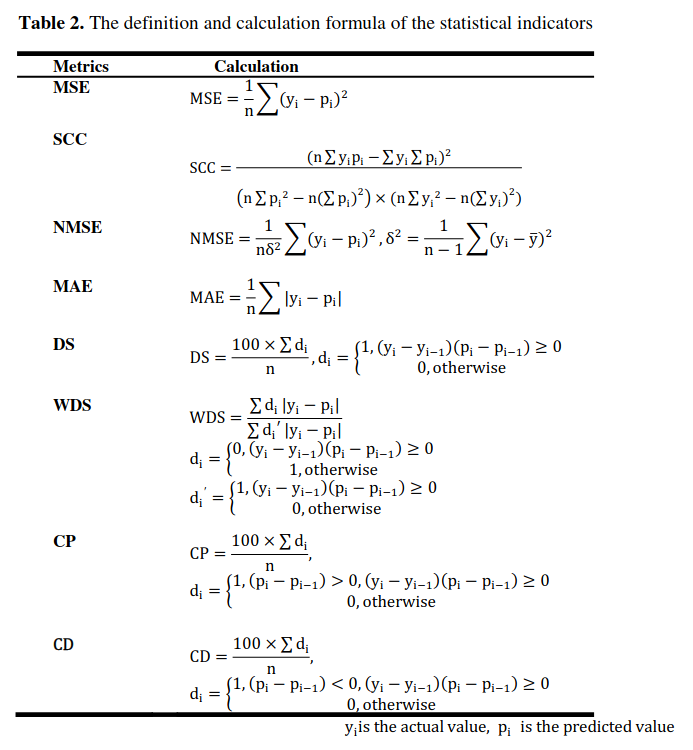
Then the data was divided into a 3 day sliding window to predict the forth day , the news data volume and price were both used in creating the time windows. Since they used time windows to predict a future day 4/5 of the data was used in the training set and 1/5 for the testing set.

To run the experiments a SVM regression function was used to examine the relationship between the news sentiment and stock prices. Libsvm which was developed by LinChih-Jen of Taiwan University was the package used and an epsilon-SVR was used to do the support vector regression.

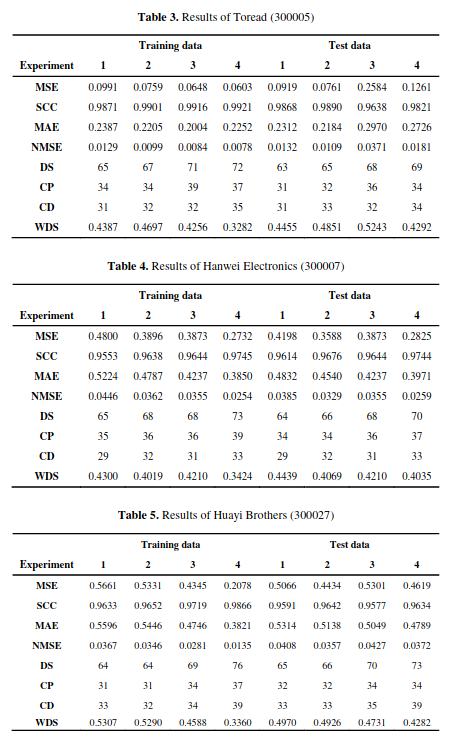
The 3 day sliding window is comprised of price at time t, t-1, and t-2 and the target was to predict the price at time t+1. Below are the 4 experiments that were run, using just price, then price and news volume, then price and sentiment value, and finally all variables.

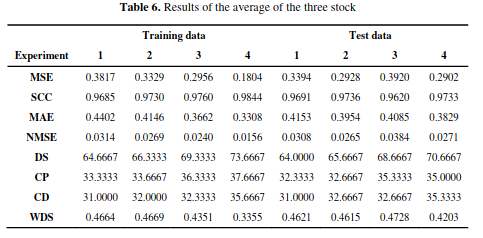


A number of metrics were used to evaluated the model: mean squared error (MSE), standardized mean squared error (NMSE), mean absolute error (MAE), multiple correlation coefficients (SCC), the direction of symmetry (DS), the weighted direction symmetry (WDS), correct upward trend (CP), and correct downward trend (CD).

­

**Results**

­



**Conclusion**

The results indicate that using even simple sentiment analysis can provide more value than simply using the volume or amount of news in forecasting and the experiments with volume and sentiment performed better than one dimensional data. There are limitations, while sentiment analysis proved useful it was done on financial news and with the growth of social media and other forms of communication it is only a piece of the puzzle that needs to be looked at. However with those other forms of communication there are new issues in parsing this data such as sarcasm that is less likely to appear in traditional financial news. The next step is to start gathering data from these sources and redo the experiments with this new data.

**Citation**

**Stock Price Forecasting with Support Vector Machines Based on Web Financial Information Sentiment Analysis,** *by: Run Cao, Xun Liang, Zhihao Ni*

Cao R., Liang X., Ni Z. (2012) Stock Price Forecasting with Support Vector Machines Based on Web Financial Information Sentiment Analysis. In: Zhou S., Zhang S.,   Karypis G. (eds) Advanced Data Mining and Applications. ADMA 2012. Lecture Notes in Computer Science, vol 7713. Springer, Berlin, Heidelberg