STOCKMAN

A System for Predicting Stock Price

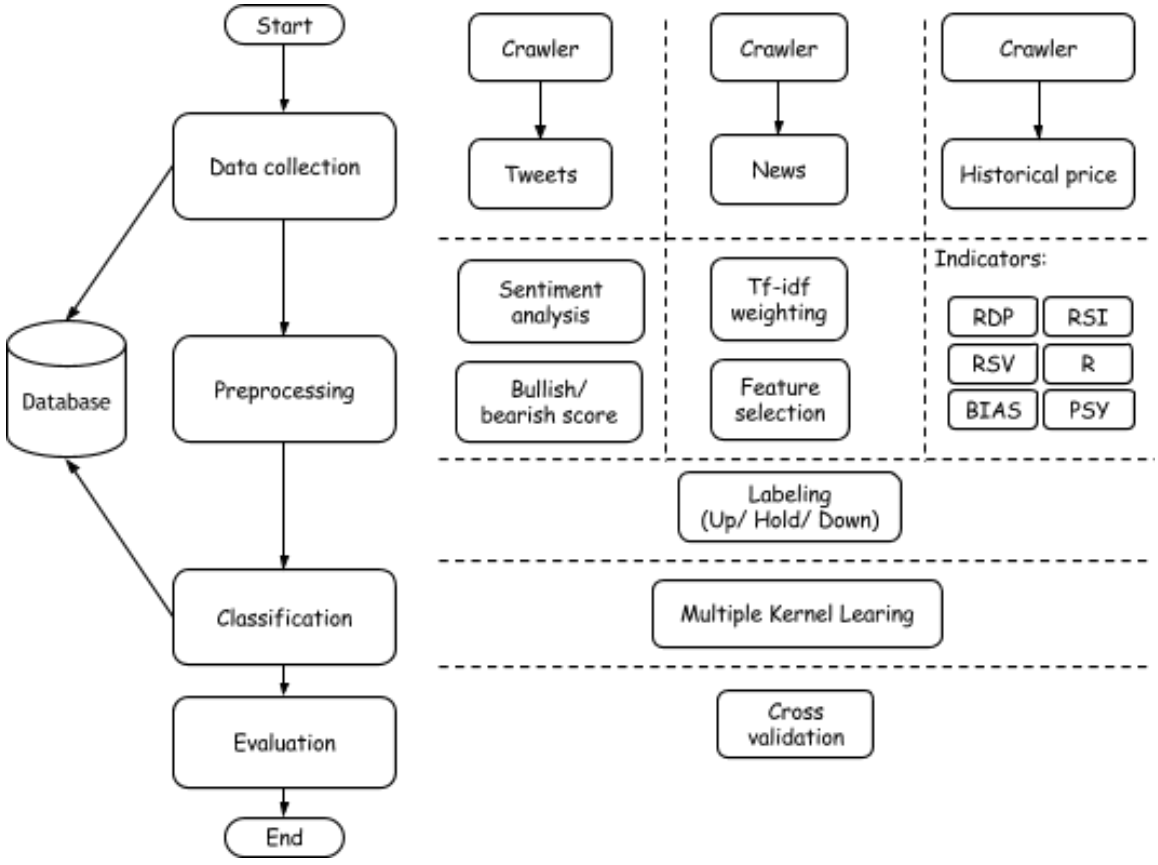
Cheuk Sio I, Chu Hoi Wai, Or Ka Wing, Yip Shing Chun
Advised by
Prof. Dik LEE

Overview

Reading various text sources to make decisions in the stock market is a simple and common method, but it is time consuming for an investor to grab and handle a lot of information. This project presents a system, Stockman, to collect financial information including tweets, news and historical price data automatically from popular websites and microblogging sites and analyze such text sources immediately. We employs the techniques from Natural language processing (NLP) and Machine learning to predict the trend of stock price.

Methodology

Three phases of the system are Data collection, Preprocessing and Classification.



(1) Data Collection

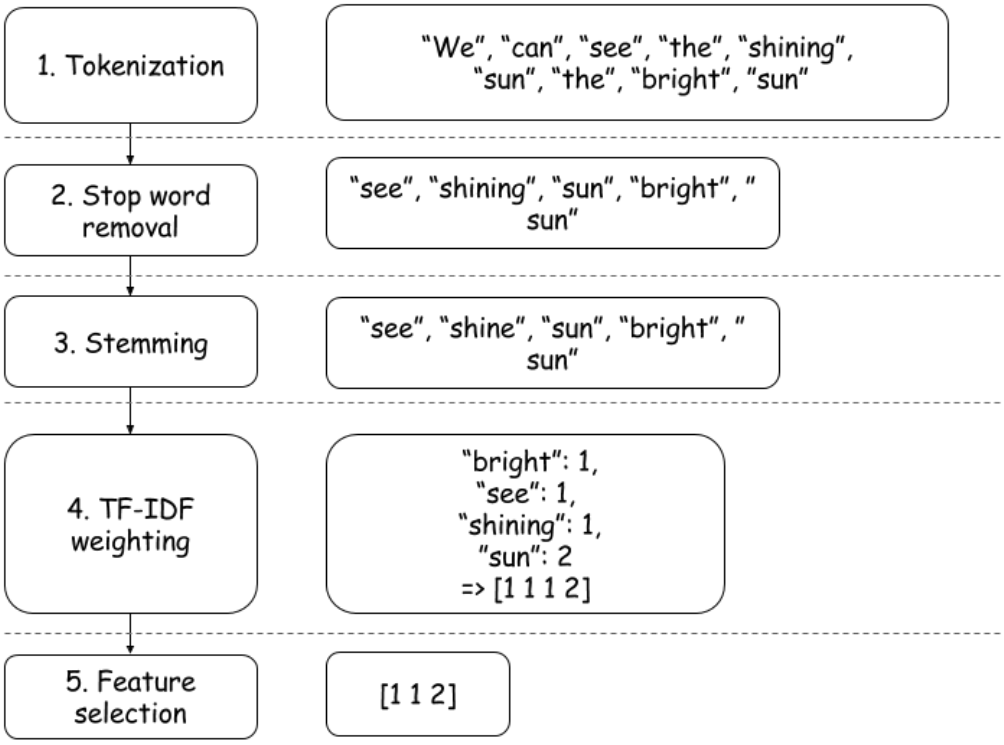
The crawler that systematically browses and downloads financial tweets from StockTwits, news from Seeking Alpha and historical prices from Google Finance.

StockTwits® Seeking Alpha α Google Finance

(2) Preprocessing

The purpose of NLP is removing useless information and hence improving the speed of processing. There are mainly five steps which include:

1. Tokenization: news articles and tweets are "tokenized" into different features
2. Stop word removal: remove the words that do not have predict value
   They includes pronouns, prepositions, articles and known vocabularies
3. Stemming: convert derived words to their stems, for example, "shining" → "shin"
4. TF-IDF weighting: TF is the frequency of words appearing. Higher, number, higher important. Some words such as "is", "am", "are" may appear frequently but do not important. IDF is used to calculate the ratio of TF in articles and TF in database to determine the importance of words.
5. Feature selection: select words that are more important



6 financial indicators are selected to improve the accuracy. These indicators are invented by the economists to predict the price changing.

- RDP: calculate the relative difference in percentage of price every 5 minutes
- RSI: shows the performance between buyers and sellers
- RSV: to predict the quick changing of stock market
- R: an reverse version of RSV
- BIAS: the value shows probability that stock price moves closer to trend line
- PSY: estimate the confidence of traders in different external environment

(3) Classification

Multiple kernel learning (MKL) is the most suitable method for learning data from different sources. It learns the predictive power of different kernels. The traditional method, Linear SVM, is also selected for classification.

Evaluation Result

There were total 2360 instances fed into the system. The result shows that tweets do not have a significant contribution to the prediction compare to news and indicators. Furthermore, the result shows that we successfully archive the objective, accuracy of 60%.

Classification result of Linear SVM:

Weight \ Minutes	5	10	15	20	25	30
Positive (%)	7.5	12.2	13.8	16.1	17.8	19.7
Neutral (%)	83.5	74.6	70.0	65.6	61.2	58.4
Negative (%)	9.0	13.2	16.1	18.3	20.9	21.9

Weight of MKL:

Weight \ Minutes	5	10	15	20	25	30
News (%)	77.6	91.9	93.2	97.1	87.9	91.9
Tweets (%)	0	0	0	0	0	0
Indicators (%)	22.4	8.1	6.8	2.9	12.1	8.1

Accuracy:

Classifier\Minutes	5	10	15	20	25	30
Linear SVM (%)	82.0±1.3	76.9±1.2	72.5±1.9	69.6±1.6	66.8±1.2	65.2±1.7
MKL (%)	61.3	63.6	65.9	66.4	65.0	66.3

From the result, we observe that the accuracy is decreasing from 5 minute to 30 minute when using Linear SVM to classify single resource. In MKL, 20 minute is the most accurate.

Conclusion

Stockman uses multiple-kernel learning to integrate news, tweets and indicators to improve prediction accuracy for stock market. It begins with crawling data from crowd sourced content platform, and afterwards filters noises by preprocessing of text such as sentiment analysis. Furthermore, a training model is built for classification. The system can predict the price movement based on this model and give the accuracy through cross validation.