Volatility Forecast in SSE&SZSE

Using Machine Learning and Sentiment Analysis

(Wang Pu, student ID：20378001)

1. Introduction

There is a huge need for effective forecasting of financial risk, which is usually implied by the related volatility. Hence the concept of financial volatility, a required parameter for pricing many kinds of financial assets and derivatives (i.e. options), is critical.

At the age of information, there is no doubt that online news can have a substantial effect on the trading price or trading volume of a stock or index, which may increase or decrease the volatility of the stock.

1.1 Case Study - China Huishan Dairy Holdings Co Ltd (06863.HK)

China Huishan Dairy Holdings Company Limited, Located at Shen Yang, Liao Ning province, produces dairy products. The Company grows and processes alfalfa and supplementary feeds, processes concentrated feed, operates dairy farms, and manufactures and sells dairy products, including milk and whey power products. China Huishan Dairy offers its products throughout China.

At 16th December 2016, Muddy Waters published a report, which said “We are short China Huishan Dairy Holdings because we believe it is worth close to Zero. We conclude Huishan is a fraud.”

Figure 1.1 The news about Huishan posted in SINA from 09 Dec to early 19 Dec.

Three days later, the part 2 was posted, which mentioned “Huishan’s reported revenue is also fraudulent. VAT data from the State Administration of Taxation show that Huishan reports a significant amount of fraudulent revenue.”

In contrast to 2 or 3 pieces of news related to Huishan Dairy per month in SINA before MW’s report, about 55 stories were released at the second half of December.

Due to MW’s short report, the price of Huishan dropped by 2.6% that day, even though the company immediately denied charges against it, and took measures to protect its stock.

What happened to this stock at March this year would not be surprised if it turns out that the report told us the truth.

One critical report may not lead to a clear trend to the price of stock, but absolutely have a substantial influence on the volatility of this stock for the next weeks, even months.

1.2 Relationship between news and volatility:

The volumes or content of news may have a profound influence on stock. Compare to volumes, the contents play a more important role in volatility forecast. Note that companies at different industry, even different companies at same industry may have different sensitivity with respect to online information.

On another hand, A big fluctuation on stock price may induce more online reports. The classification that splits the news to the report before the event and the report after the event, as well as other classification, should be considered in a complete model.

Period	06 -15	16-30
Close price	2.50%	6.78%
Volume	5.16	37.33

Table 1.1 The price's and volume's volatility of stock 06863.HK at period before and after MW's report.

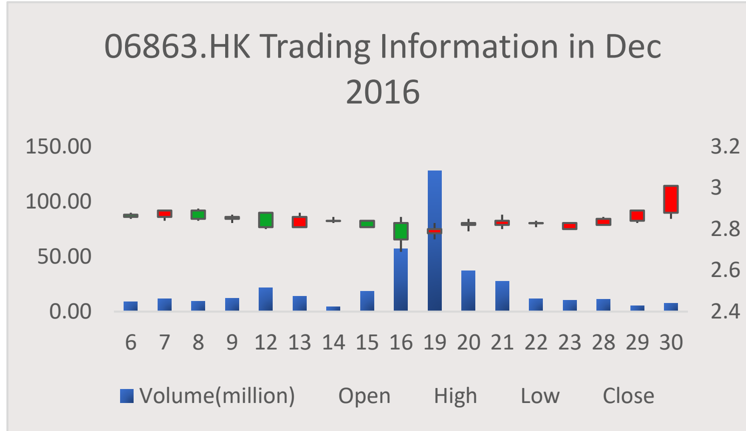

Figure 1.2 The historical market data for stock 06863.HK from 6 Dec 2016 to 30 Dec 2016.

2. News crawler

I crawled all the news webpages published at year 2016 under *个股新闻* at www.finance.qq.com, and stored them at a SQLite database. (For the details about technology for crawler, see the codes)

id	news_id	source_id	pubdate	time_window	url	title	content	sentiment
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	stock-201612...	1	2016-12-29 0...	NULL	http://stock.q...	曾经定增玩得...	时报君今...	0.0
2	kuaibao-2016...	2	2016-12-28 1...	NULL	http://finance...	"新华—浦发"...	晶报讯 浦...	2.0
3	stock-201612...	3	2016-12-28 0...	NULL	http://stock.q...	石化板块有资...	昨日沪深两市...	-5.0
4	finance-2016...	4	2016-12-26 1...	NULL	http://finance...	银行业周报:...	一周消息...	2.0
5	finance-2016...	5	2016-12-26 1...	NULL	http://finance...	银行行业周报...	报告要点 ...	-14.0
6	stock-201612...	6	2016-12-24 0...	NULL	http://stock.q...	2016年A股大...	编者按: 再过5...	31.0
7	stock-201612...	7	2016-12-22 0...	NULL	http://stock.q...	浦发银行副行...	今年以来, 银...	9.0
8	stock-201612...	1	2016-12-20 1...	NULL	http://stock.q...	浦发银行: 非...	浦发银行(600...	0.0
9	finance-2016...	8	2016-12-20 1...	NULL	http://finance...	2017年策略...	2017年金...	18.0
10	stock-201612...	9	2016-12-20 0...	NULL	http://stock.q...	光大、浦发、...	中国证券...	-29.0
11	finance-2016...	4	2016-12-19 1...	NULL	http://finance...	银行行业周报...	一周消息...	2.0
12	finance-2016...	10	2016-12-19 1...	NULL	http://finance...	银行行业周报...	板块表现...	11.0
13	stock-201612...	1	2016-12-16 2...	NULL	http://stock.q...	浦发银行: 获...	浦发银行(600...	1.0
14	finance-2016...	4	2016-12-15 1...	NULL	http://finance...	银行行业201...	核心观点 ...	7.0
15	finance-2016...	11	2016-12-13 0...	NULL	http://finance...	银行行业投资...	事件: 经...	31.0

Figure 2.1 A snippet of elements of Table "News" at the Data Base, named "VF.sqlite".

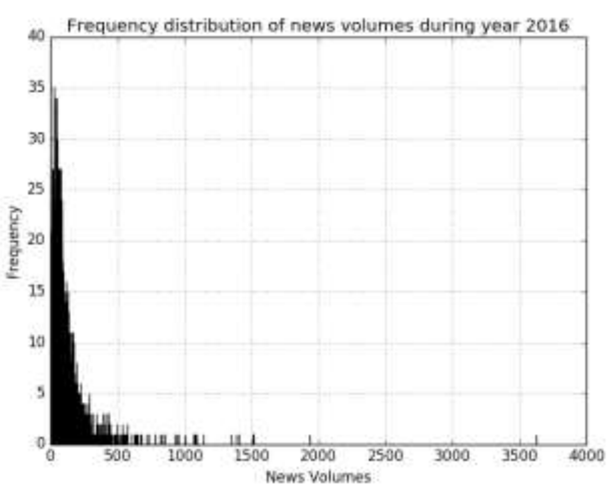

Figure 2.2 Frequency Distribution of News Volume

# of companies	3,125
# of Sources	960
News volume	334,946
mean	107
Standard deviation	144

Table 2.1 Statistic of news volume

MIN	25%	50%	75%	MAX
0	36	71	130	3627

Table 2.2 Frequency Distribution of News Volumes

3. Sentiment Analysis

The sentiment of a news chapter is based upon the sentiment values for all its keywords, that is those containing emotional polarity.

Word sets	Description
POSITIVE	A list of Chinese words that have positive emotional polarity, which includes a set of 2810 words. (released by NTUSD)
NEGATIVE	A list of Chinese words that have negative emotional polarity, which includes a set of 8276 words. (released by NTUSD)
PRIVATIVE	A list of privative Chinese words, which includes 11 words (不,不是,没,没有,非,并非,无,未,未能,难,不见得)
The following five sets are modifiers, whose intensities decrease while i increases.(released by HowNet)	
MODIFIER1	69 modifier words, with WEIGHT1 = 2
MODIFIER2	42 modifier words, with WEIGHT2 = 1.8
MODIFIER3	37 modifier words, with WEIGHT3 = 1.6
MODIFIER4	29 modifier words, with WEIGHT4 = 1.4
MODIFIER5	12 modifier words, with WEIGHT5 = 0.8

Table 3.1 Word Sets used in calculating the keyword sentiment

I calculated keyword sentiment by counting matches with the given sentiment dictionary, using the algorithm proposed by Desheng Dash Wu and David L. Olson (2015). The sentiment for an entire chapter is calculated by summing the sentiments for all the keywords contained in that chapter.

I incorporate the sum of absolute sentiment value into the current forecasting model, and predict how volatility will move in the immediate future using a more comprehensive perspective.

Note that I segmented the news' text with *jieba*, *one of* the best Python Chinese word segmentation modules. The dictionaries of NTUSD and HowNet are released at 2007, and updated version are NOT available online.

News ID	Company code	News title	News body	News sentiment	Time window
stock-2016122000825 2	SH600000 浦发银行	光大、浦发、招行成承销商中前三大债券"踩雷王"	中国证券网讯 根据同花顺统计，仅12月便已有3起债券违约，至此，年内债券违约事件达到63起， 涉及金额高达378.94亿元，是2015年违约金额的三倍多。。。。。	-29.0	2016-12-18
finance-20160826025468	SZ000001 平安银行	亮眼成绩破"寒冬"之说 零售业务成银行转型重点	"2016年上半年，平安银行资产总额28009.8亿元，较年初增长11.7%；营业收入547.69亿元，同比增长17.59%；准备前营业利润361.56亿元、同比增长28.26%；实现净利润122.92亿元、同比增长6.10%......"	55.0	2016-07-31

Table 3.1 A snippet of news entries

4. Volatility Forecast

Given the data base that I have crawled, I selected the data of stocks that were listed on the exchange before 01/01/2016. then resampled the data to week bins by calculating the mean and variance of price daily return, and summing the absolute value of every news chapter's sentiment value during every week. Finally, I got the data used to train and forecast volatility.

Here are some details in the "total.csv":

- "date": label of time window; there are 50 weeks in year 2016, excluding the first week of Feb and the first week of Oct due to holiday in China;
- "code": the code of stock;
- "p_var": variance of price daily return within this time window
- "mean_return": average of price daily return;
- "num_news": the news volumes within the next week from the date;
- "sum_abs_sent": sum of absolute value of every financial story released within the next week.

Volatility refers to the standard deviation or variance of the change in value of a financial instrument within a specific time span. The GARCH system is widely employed in modeling financial time series that exhibit time-varying volatility clustering. In this section, we develop a GARCH system by incorporating financial information into the usual framework.

The GARCH model, proposed by Bollerslev in 1986, 2 can be formulated as$y_t = \mu_t + \varepsilon_t$

- $\varepsilon_t | \psi_{t-1} \sim N(0, \sigma_t^2)$
- $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j}^2$

where, $\alpha_0 > 0, \alpha_i, \beta_j > 0$; p, q represents the time lags.

The GARCH model uses ε_t as a function of those exogenous inputs, which have some affect on financial volatility. The GARCH model bases its conditional distribution on the information set available at time t. Freisleben and Ripper point out that the parameter β_j in Equation (5.3) describes the stock return's immediate reaction to new events in the market, mostly in the form of financial news. Meanwhile, the fast development of the internet enables us to acquire the online financial information in a real-time, exhaustive fashion. Considering these factors, designating financial information sentiment value as one variate of ε_t is justifiable.

The idea is to formulize ε_t using y_t and W_t, where W_t is the sentiment value of on-line financial information on time t. For simplicity, assume that ε_t^2 is a linear combination of nonlinear function of y_t^2 and W_t^2, that is

- $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \chi_{t-i}(\sigma_{t-i}^2) + \sum_{j=1}^{q} \beta_j \varphi_{t-j}(y_{t-j}^2) + \sum_{k=1}^{q} \gamma_k \phi_{t-k}(W_{t-k}^2)$

where $\chi_{t-i}, \varphi_{t-j}, \phi_{t-k}$ represent the undetermined nonlinear correlations.

I used SVM and Random Forests to dynamic train and forecast the volatility of stocks. For SVM, the kernel function used is RBF, and the penalty parameter c and the RBF kernel parameter g are set as c = 64 and g = 1/3. For Random Forests, I set M=1 and chose full-grown trees.

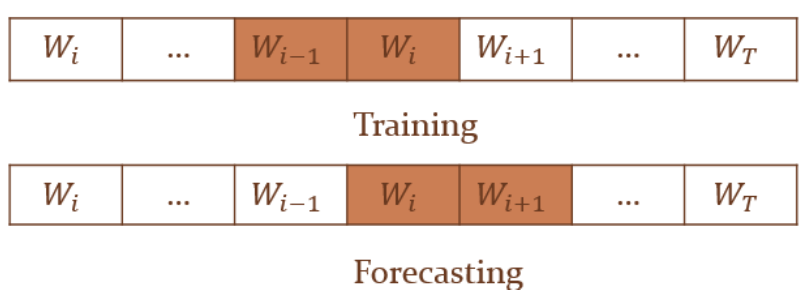

Table 4.2 Sliding time window learning and forecasting

Besides, two major performance metrics are introduced in this experiment to evaluate the aggregated forecast performance for all the 3,125 companies: adjusted squared correlation coefficient (ASCC) and volatility trend forecast accuracy (VTFA). ASCC and VTFA are computed based on the forecasting values for each time window.

- The *adjusted squared correlation coefficient* evaluates the correlation of all the explanatory variables to the response variable. The closer this value is to 1, the better regression result is achieved.
- The *volatility trend forecast accuracy* is the proportion of companies with an accurately predicted volatility trend among all the companies.

Table 4.3 Forecast results for 2500+ listed companies during the year 2016

forecast output	forecast input	train set	predict set	SCC(R^2) SVM	SCC(R^2) RF	VTFA SVM	VTFA RF
1/17/2016	1/10/2016	2533	2531	0.705	0.845	62.07%	73.84%
1/24/2016	1/17/2016	2531	2545	0.840	0.832	57.56%	39.49%
1/31/2016	1/24/2016	2545	2539	0.795	0.846	43.68%	30.09%
2/14/2016	1/31/2016	2539	2536	0.899	0.842	68.22%	69.99%
2/21/2016	2/14/2016	2536	2526	0.759	0.810	63.14%	69.52%
2/28/2016	2/21/2016	2526	2518	0.662	0.810	62.03%	63.03%
3/6/2016	2/28/2016	2518	2514	0.832	0.855	57.76%	56.96%
3/13/2016	3/6/2016	2514	2516	0.759	0.805	72.38%	54.49%
3/20/2016	3/13/2016	2516	2518	0.761	0.816	55.16%	50.48%
3/27/2016	3/20/2016	2518	2495	0.815	0.818	70.38%	69.30%
4/3/2016	3/27/2016	2495	2473	0.685	0.810	58.84%	57.50%
4/10/2016	4/3/2016	2473	2460	0.652	0.816	62.11%	57.03%
4/17/2016	4/10/2016	2460	2463	0.676	0.812	64.51%	68.94%
4/24/2016	4/17/2016	2463	2476	0.674	0.809	43.58%	36.79%
5/1/2016	4/24/2016	2476	2483	0.766	0.802	68.55%	67.10%
5/8/2016	5/1/2016	2483	2483	0.611	0.811	58.16%	51.39%
5/15/2016	5/8/2016	2483	2487	0.716	0.806	61.12%	61.32%
5/22/2016	5/15/2016	2487	2483	0.793	0.841	64.60%	59.52%
5/29/2016	5/22/2016	2483	2481	0.731	0.840	53.45%	64.77%
6/5/2016	5/29/2016	2481	2486	0.617	0.806	43.16%	40.35%
6/12/2016	6/5/2016	2486	2489	0.543	0.835	56.93%	67.82%
6/19/2016	6/12/2016	2489	2502	0.618	0.841	37.85%	33.61%
6/26/2016	6/19/2016	2502	2508	0.727	0.831	63.44%	63.28%
7/3/2016	6/26/2016	2508	2525	0.640	0.828	68.20%	65.47%
7/10/2016	7/3/2016	2525	2525	0.613	0.807	67.45%	65.39%
7/17/2016	7/10/2016	2525	2535	0.639	0.792	64.18%	63.12%
7/24/2016	7/17/2016	2535	2545	0.576	0.811	63.42%	72.18%
7/31/2016	7/24/2016	2545	2556	0.593	0.816	46.01%	37.17%
8/7/2016	7/31/2016	2556	2569	0.770	0.820	66.29%	66.60%
8/14/2016	8/7/2016	2569	2577	0.579	0.828	62.63%	61.82%
8/21/2016	8/14/2016	2577	2583	0.568	0.793	67.44%	65.51%
8/28/2016	8/21/2016	2583	2576	0.388	0.831	64.95%	58.54%
9/4/2016	8/28/2016	2576	2581	0.502	0.827	64.20%	66.18%
9/11/2016	9/4/2016	2581	2561	0.556	0.811	67.32%	73.99%
9/18/2016	9/11/2016	2561	2547	0.625	0.820	47.39%	39.81%

forecast output	forecast input	train set	predict set	SCC(R^2) SVM	SCC(R^2) RF	VTFA SVM	VTFA RF
9/25/2016	9/18/2016	2547	2560	0.696	0.803	67.62%	73.20%
10/9/2016	9/25/2016	2560	2556	0.570	0.818	58.84%	57.16%
10/16/2016	10/9/2016	2556	2577	0.613	0.808	66.08%	64.88%
10/23/2016	10/16/2016	2577	2582	0.549	0.804	64.06%	59.76%
10/30/2016	10/23/2016	2582	2588	0.517	0.826	62.17%	62.87%
11/6/2016	10/30/2016	2588	2590	0.728	0.838	70.19%	69.54%
11/13/2016	11/6/2016	2590	2593	0.609	0.838	57.50%	54.69%
11/20/2016	11/13/2016	2593	2578	0.526	0.806	63.42%	64.93%
11/27/2016	11/20/2016	2578	2573	0.702	0.832	69.69%	71.08%
12/4/2016	11/27/2016	2573	2586	0.554	0.800	57.12%	52.13%
12/11/2016	12/4/2016	2586	2590	0.634	0.832	51.97%	63.47%
12/18/2016	12/11/2016	2590	2584	0.497	0.811	39.94%	36.73%
12/25/2016	12/18/2016	2584	2592	0.690	0.828	62.11%	60.57%

Table 4.3 presents a demonstration of the asset price volatility for 2,500+ companies using information sentiment during the year 2016.

On average, both methods can achieve nearly 60% of volatility trend forecast accuracy (60.18% for SVM and 59.02% for RF), while the variance of VTFA for Random Forests are twice it for SVM. An average of 65.77% of adjusted R squared was achieved for SVM, and 82.02% for RF, giving convincing evidence to the correlations between these factors.

Furthermore, Figure 4.1-4.4 illustrate the price volatility forecasts for two specific companies out of the 3,125. it can be found that the predicted values, under most circumstances, correspond well to the actual values for the first stocks, although for occasional huge oscillations the forecast result is not very good. while for stocks that has small volatility for the entire time, the predicted values are pretty higher than the actual values.

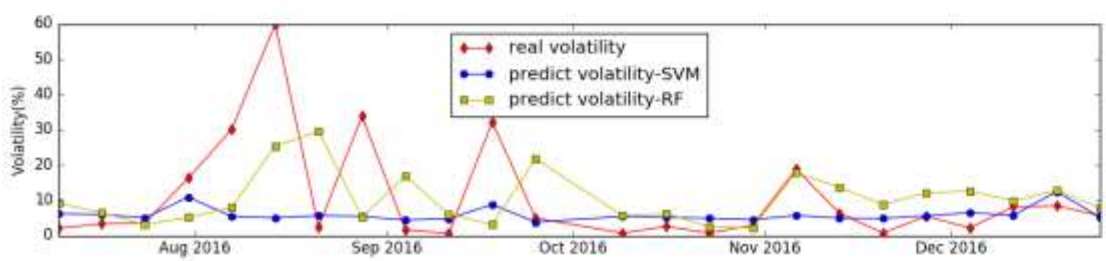

Figure 4.1 Price volatility forecast result for company VanKe (SZ000002) over all the time windows

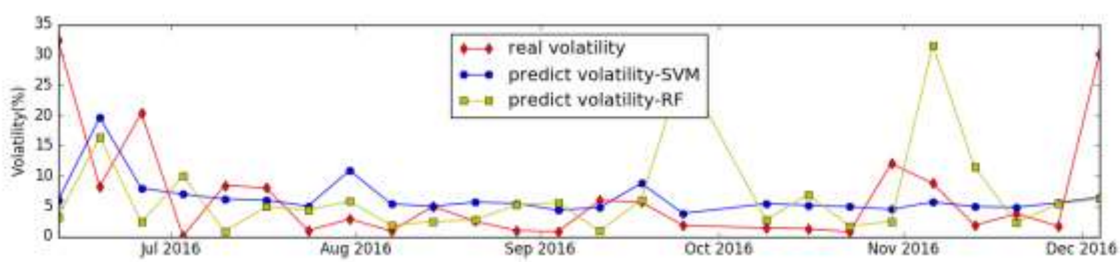

Figure 4.2 Price volatility forecast result for company LeTV (SZ300104) over all the time windows

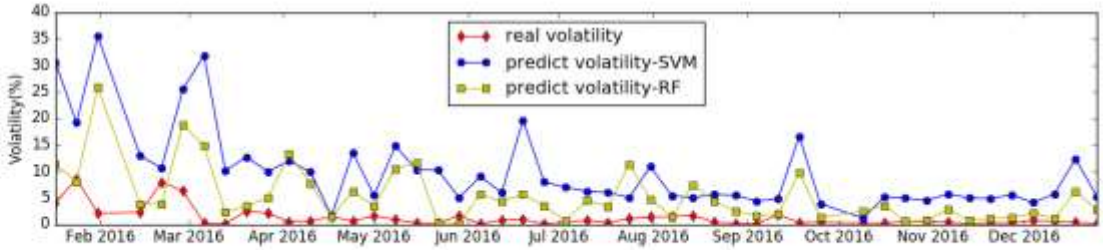

Figure 4.3 Price volatility forecast result for company Ping An Bank (SZ000001) over all the time windows

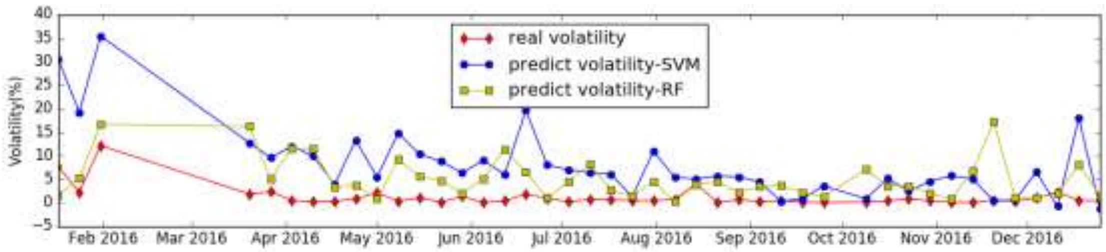

Figure 4.3 Price volatility forecast result for company Shanghai Pudong Dev Bank (SH600000) over all the time windows

5. Conclusions and Further Study

I have introduced GARCH-based SVM and Random Forests to investigate the correlations between asset price volatility and information sentiment. Both methods are capable of achieving favorable prediction results; SVM performs better in predicting the volatility trend than RF, since it has less variance.

The empirical studies can be useful to financial investors, portfolio holders, academicians, etc. in the sense that they provide an alternative tool to forecast volatility and trend.

As for as I see right now, I can improve the forecast ability of this model by two areas:

1) Chinese Financial Sentiment Dictionary: As mentioned before, the dictionary available was released many years ago, and it was created for general text analysis. To achieve better sentiment analysis of financial reports, we need a new special sentiment dictionary.
2) Multi-Evaluation Factors of Financial News: Considering other features, such as the classification of news, source of news, etc. may give a better insight to financial news. This waits to discuss.