ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

Advances in Decision Sciences

Volume 25
Issue 2
June 2021

Michael McAleer

Editor-in-Chief
University Chair Professor
Asia University, Taiwan

Published by Asia University, Taiwan

ADS@ASIAUNIVERSITY

Social Media and Analytics: A Case Study of Tweets Data *

Hak J. Kim **

Department of Information Systems and Business Analytics
Hofstra University

Revised: May 2021

* The author thanks the reviewer for very helpful comments and suggestions.

** Correspondence: hak.j.kim@hofstra.edu

Abstract

Social media is emerging as a main communication channel in business and society. People use it in their daily life to share information with their friends and family. In this paper, we attempt to analyze Tweets data for understanding the relationship of social media and stock market as a case study. Data is collected more than 1.6 million tweets and RapidMiner software is used for their analysis. Our result shows that social media has relation with stock mark. It means that social media has an impact on financial market. However, there are some limitations to use social media for better decision making in investment. In a future paper, we need to develop more sophisticated analytics model and integrate different types of data to produce better results.

Keywords: Social Media, Big Data, Analytics, Twitter, Financial Stock Price.

JEL: C50, C90, C59, M15, M20.

1. Introduction

Today's computing paradigm change is very much to some of the other shifts that have happened in information and communications technology (ICT) industry, such as mainframe to client/server and to cloud (Kuo, 2011). It is very different from previous shifts because of emerging multiple innovative technologies including mobile computing, cloud computing and social media technologies, . A bunch of different activities are happening whether it is social, whether it is mobility, whether it is cloud, and just as insatiable demand for information. It is like a huge storm or Tsunami [2].

As mobile cloud computing (Fernando et al., 2013) is taking over the position of the Internet, people can use their smartphones more often to communicate with family, friends, and other people regardless of time and place. Riche applications and faster services (i.e., 4G or 5G technologies) have brought explosive growth of social media market including YouTube, Twitter, Instagram, and Facebook.

We are certainly living in the era of social media. Social media today is popularized as new paradigm of communication. Unlike traditional structured data, social media data is unstructured like text and audio/video clips. In the past, those data could not be measured and analyzed because of lack of technologies and tools for collecting and analysing them.

Recent advancement of computing technologies and data science, the unstructured data can be collected and analysed. Social media data is a typical type of these unstructured data which is called 'Big Data'(Manyika et al., 2011). Social media analytics is simple a set of methodologies and techniques to collect and analyse social media data for producing meaningful information or insights.

Previous studies (Gruhl et al., 2005; Liu et al, 2007; Choi and Varian, 2009; Asur and Huberman, 2010) have been done related to social media. For instance, Karabulut (2011) analyzed Facebook's Gross National Happiness (GNH) index He concluded that GNH could predict daily returns in the US equity market. This paper builds upon keywords from tweets to create its own index rather than to use a published index.

The paper presents to the understanding of social media analytics and applies it to the financial sector focused on a stock market as a case study. The purpose of this paper is to analyze social media data (that is, tweets) and identify its relationship with stock market data (that is, Dow stock price index). The assumption is that the keywords represent investor sentiment. This paper assumes that that an investor's sentiment plays an important role for making decision when they buy and sell stocks. As a result, tweets can affect the daily traction in the stock market. In the future, the analytics model will be built to predict the stock movement using any social media data.

This paper is organized as follows. The second section is briefly presents research methods and overviews data. A simulation model using RapidMiner is explained in the third section. The fourth section discusses simulation results and its findings. Finally, a conclusion and future paper are presented in the fifth section.

2. Research Methods and Data

2.1 Hypotheses

The paper establishes the following hypotheses:

- **Hypothesis 1:** There exists a positive correlation between real return and estimated return.
- **Hypothesis 2:** There is high fluctuation in real return and estimated return in the earlier periods from 2009 (financial crisis), whereas there is low fluctuation in the later periods from 2009.
- **Hypothesis 3:** Seemingly good words (i.e., good, up, high, buy…) will be associated with high returns, whereas bad words (i.e., bad, down, low, sell…) will be associated with low returns.

2.2 Data Collection and Cleaning

The paper is based on the assumption that tweets reveal an investor's sentiment and thus provide a basis for predicting financial market. We collected over one million tweets data from 2006 to 2011. The primary data was the text file which has 1,608,621 rows representing each individual tweet. The data includes date and time, twitter ID, and keywords. Figure 1 shows a sample of raw tweets data.

[Figure 1 here]

The data is broken into three groups equally: 536,207 tweets per group). We process data cleaning to each group's data set. First, a tweet data which is missing a keyword is removed because it presents no value to the analysis. Second, the date and time is shortened to include only the date and removed time data because we calculate only daily return (not time basis). Twitter ID data is also deleted because it is irrelevant for our data analysis. After data cleaning, data size are decreased from 1,608,621 rows and 4 columns (fields) to 1,130,577 rows and 2 columns (Table 2).

A new column is created for adding 'Return' data. Return is calculated by using the following single period arithmetic formula:

$$R = \frac{P_1 - P_2}{P_1}$$

where P_1 represents the adjusted close price of the previous day.

P_2 represents the adjusted close price of the present day.

The data for the daily adjusted close price are taken from the S&P500 index in Yahoo! Finance (2021).

2.3 Modelling and Simulation

The tool for tweets data analysis is RapidMiner (2021). It is a GUI mode open-source system with more than 500 operators for data mining and data analysis. The main components of RapidMiner are operator, process, and repository. Operator is simply a task to be processed. Processor is connecting between operators. Repository is a kind of storage that retrieve and save data.

The first step is to convert the raw data into a format readable for RapidMiner. After the data preparation stage, the next step is to find the return associated with each keyword. For example, the keyword "good" may be associated with positive returns, whereas the keyword "bad" may be

associated with negative returns. Then, the next step is to develop an estimated (hypothetical) return schedule based on the returns associated with each keyword. The final step is to analyze the relationship between the estimated return schedule and the real return schedule. The hypothesis is that the estimated return schedule and the real return schedule will have a strong positive correlation.

Figure 2 shows our research model for simulation. The first process is the "Keyword" process which is to produce return relevant keywords. The three CSV files were imported using three "Read CSV" operators. The "Append" operator combined all three sets of data into a single set of data. "Select Attributes" operator selected only the relevant data columns to be used for the analysis.

[Figure 2 here]

Since the relevant columns are "Keyword" and "Return(date)", the "Aggregate" operator grouped the keywords and computed the average of Return(date). The "Write Excel" operator enabled the final results to be recorded in an Excel file. Return (keyword) column is the return associated with each keyword. It also has a "frequency" column which shows how many times a word appears in tweets. Low frequency shows that the return associated with the keyword is unreliable because of small sample size. As a rule of thumb, words with a frequency of less than 100 were deleted for increased accuracy in average return. For creating a process that yields a hypothetical or estimated return schedule based on the keywords, we create files for input. As shown in Figure 3, a column "Return(date)" is added to the original file using the LOOKUP function.

[Figure 3 here]

The file has columns to represent the day's return. Since there are multiple rows with the same date, a process is needed to perform aggregation. The process "Value" is able to group the rows with the same date, and compute the average of all the returns in that group. It is very similar to the "Keyword" process, but the only difference is that the "Value" process groups by date, whereas the "Keyword" process groups by keyword. Returns with a low sample size had to be deleted. Therefore, as a rule of thumb, any dates with less than 200 rows are deleted to improve accuracy. As shown Figure 4, it contains columns that represent the estimated returns corresponding to each date.

[Figure 4 here]

Another column, “Return(date)”, is added As shown in Figure 5. The column “Return(date)” represents the real return of that day, whereas “Estimated Return” represents the estimated or hypothetical return. This allowed for a comparison between the estimated return and the real return.

[Table 5 here]

A simple process called “CM” was used to find out the correlation between the estimated return and the real return. As shown in Figure 6, the process has three operators: “Read Excel”, “Select Attributes” and “Correlation Matrix”. The “Read Excel” operator takes in the Excel file. Then, the “Select Attributes” operator enables the selection of attributes that are relevant to this process. Finally, the “Correlation Matrix” computes the correlation between all the selected attributes. Instead of writing the result in Excel format, this process shows the results within RapidMiner.

[Figure 6 here]

4. Results Analysis

The analysis revealed some interesting results in the process. However, the analysis did not result in a useful model that could predict the future stock market changes due to many reasons.

Hypothesis 1: The hypothesis is rejected. The correlation matrix revealed a weak positive correlation between the estimated return and real return. As shown in Figure 7, the correlation between EC and RC was 19.2%.

[Figure 7 here]

The reason behind the low correlation may be the huge chunk of data loss from dates November 2009 to January 2010. The cause could not be identified. However, there are some possible explanations. Switching the file formats back and forth may have caused this loss. Accidental deletion in the process of handling data manually could be a cause. Using many different processes instead of using a combined single process in RapidMiner may have caused data loss (Figure 8)

[Figure 8 here]

Hypothesis 2: The hypothesis is accepted. In early periods of the data, real return had high fluctuations in daily price changes. However, estimated return did not experience high fluctuations in the early periods. The rationale behind the anticipated high fluctuation is the financial crisis of 2007-2008. The available data ranges from October 2006 to March 2010. This includes the 2007-2008 financial crisis period where daily changes can be extreme. As shown in Figure 9, the graph shows that there are high fluctuations in daily changes in the earlier periods close to the crisis and low fluctuations in the later periods.

[Figure 9 here]

However, the graph in Figure 10 does not present high fluctuations in the earlier periods.

[Figure 10 here]

Hypothesis 3: The hypothesis is partially accepted. The hypothesis is that seemingly good words were associated with high returns, whereas bad words were associated with low returns.

In Figure 11, the word associated with the highest returns was “acquiring”. Interestingly, “acquired” and “acquires” had the third highest return and the fifth highest return respectively. Also, “merge” and “merger” had the eighth and ninth highest return. It is unclear why words related to mergers and acquisitions are associated with high return in the stock market. Other words associated with high returns include: “higher”, “upside”, “gain”, “up”, “profit”… , and so on. The word associated with the

lowest return was “downside”. Words that are characterized by negative returns include: “down”, “loss”, “bottom”, “low”, “bad”…, and so on.

[Figure 11 here]

5. Conclusion

5.1 Managerial Implications

Meaning 1: We attempt to find the value of social media data through empirical data.

Meaning 2: Using the real social media data, we attempt to predict the future financial market.

5.2 Limitations

Limitation 1: No previous studies

There were many difficulties and problems associated with this analysis. The biggest difficulty was the availability of resources – the lack of previous studies on the subject forced the use of unproven methods and models. For instance, simple arithmetic mean was used when computing the returns associated with the keywords. Outliers may have had a big impact on skewing the arithmetic mean. Using other measures of central tendencies, such as median, may have helped increase the accuracy of the returns.

Limitation 2: Reality Issue

Data is not directly collected from the real-world. We got it from the vendor. There is an issue of reliability. It is impossible to get real data without the vendor support.

Limitation 3: Low performance system

The analysis was done from a laptop computer with subpar processer and low RAM. RapidMiner would frequently lag and freeze when working with large amounts of data. Working in such conditions increased the time spent on importing data, converting format and producing results. The analysis would have gone much faster with a better computer, allowing additional analysis on the subject

In addition, through this opportunity, we were able to learn about RapidMiner and apply it in social media data. We saw much potential in RapidMiner as an analytics tool, and were constantly amazed by what it could accomplish. More analytical operators and functions are being added to RapidMiner, so that it presents opportunities for further research on this topic.

References

Asur, S., and Huberman, B. (2010), Predicting the Future with Social Media arXiv:1003.5699v1

Choi, H & Varian, H. (2009), Predicting the present with google trends, *Technical report*, Google.

Fernando, F., Loke, S., and Rahayu, W. (2013), Mobile cloud computing: A survey, *Future Generation Computer Systems*, 29(1), 84-106.

Gruhl, D, Guha, R, Kumar, R, Novak, J, & Tomkins, A. (2005), The predictive power of online chatter, *ACM*, 78-87.

Kuo, C. (2011), Paradigm Shifts in Modern ICT Era and Future Trends, *In Proceedings of the 10th International Symposium on Signals*, Circuits and Systems (ISSCS).

Liu, Y, Huang, X, An, A, & Yu, X. (2007), ARSA: a sentiment-aware model for predicting sales performance using blogs, *ACM*, 607-614.

Manyika, G., Chui, J., Brown, M. Bughin, B., Doobs, J., Roxburgh, R., and Byers, A. (2011), Big Data: The Next Frontier for Innovation, Competition, and Productivity, *Report*, McKinsey Global Institute.

RapidMiner (2021), https://rapidminer.com/

Wegmuller, M., Weid, J., Oberson, P., and Gisin, N. (2000), High resolution fiber distributed measurements with coherent OFDR, *in Proceedings ECOC'00*, paper 11.3.4, 109.

Yahoo!Finance (2021) https://finance.yahoo.com/

Zhang, S., Zhu, C., Sin, J., and Mok, P. (1999), A novel ultrathin elevated channel low-temperature poly-Si TFT, *IEEE Electron Device Lett.*, 20, 569–571.

Figure 1

Raw Tweets Data

File Edit Format View Help

```
20070306212211  $$$     0005887328
20070306221120  $$      0005888593
20070307155450  $wife   0005905967      up
20070309014235  $$      0005947467      bull,bought
20070309060022  $ney    0005959151      soon
20070309152822  $boss   0006048741
20070309194559  $boss   0006122021      today
20070312193229  $...    0007093051
20070313011634  $...    0007220691
20070313175359  $$$     0007466691
20070314220940  $^      0007944461
20070315042223  $k      0008053671
20070316194743  $$      0008698881      today,up
20070317025538  $$      0008823181      bad
20070318024912  $$$     0009181191
20070318231503  $$      0009438481      smart,buy,smart
20070319012838  $.      0009479741      year
20070319184415  $gruber_tweets  0009758071
20070319184415  $gruber_tweets  0009758071
20070320030941  $haun   0009898901      close
20070320065312  $prog   0009958241
20070321185826  $$      0010625571
20070322045642  $kin    0010868201
20070322211343  $$$$$   0011197911
20070323035817  $$$     0011359891      update
20070323223522  $tonic. 0011785361
20070323230843  $prefix 0011799091
20070324223325  $$$$    0012219241      worth
20070325203510  $food   0012609141
20070325203510  $inboxSize      0012609141
20070325203510  $sleep  0012609141
20070325203510  $water  0012609141
20070326003606  $$      0012710071
20070326132542  $hit    0012972991      hate
20070326132634  $hit... 0012973531
20070326173024  $oft:   0013084671      made
20070327050826  $.      0013446471
20070327215256  $$$     0013881081
20070327231511  $WORKSFORME     0013919521
```

Figure 2

A Simulation Model with RapidMiner

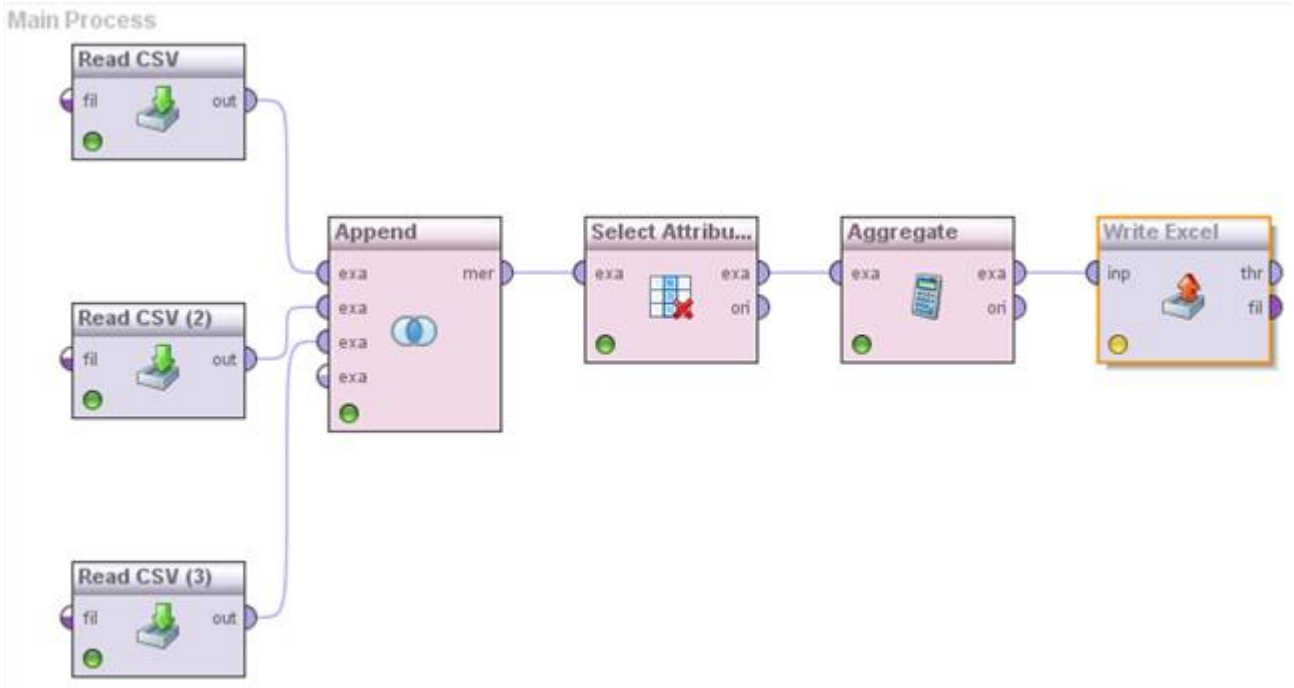

Figure 3

Result of Keyword Analysis

Keyword	Return(keyword)	Frequency
acquire	0.00131280	371
acquired	0.00267631	293
acquires	0.00255162	214
acquiring	0.00319691	134
action	0.00104051	3573
alpha	-0.00097745	645
announce	0.00054675	667
announced	0.00076710	1355
announcement	0.00137009	619
announces	-0.00049764	945
announcing	-0.00064544	176
bad	-0.00006385	11646
bear	-0.00008253	2193
beta	0.00168895	456
bond	0.00036867	655
bottom	-0.00037005	2224
bought	0.00075457	17442
bounce	-0.00052624	2186
breaking	0.00057719	2180
breakout	0.00266934	2846
bull	0.00038834	3649
buy	0.00060098	22565
buying	0.00052332	7078
call	0.00062581	11579
cap	0.00023919	1351
cash	-0.00003897	6921
charges	0.00089415	632

Figure 4

Estimated Return

	A	B	C
1	**Date**	**Estimated Return**	**Keyword Per Day**
2	09/15/2008	0.0002798024093023	215.00
3	10/06/2008	0.0001025773080189	212.00
4	10/10/2008	0.0001732615737705	244.00
5	10/16/2008	0.0002552908557377	244.00
6	10/17/2008	0.0004193588322581	220.00
7	10/21/2008	0.0003545098779167	241.00
8	10/22/2008	0.0002512865661905	210.00
9	10/23/2008	0.0002895931437209	215.00
10	10/24/2008	0.0001388618072519	262.00
11	10/27/2008	0.0003592420009615	209.00
12	10/28/2008	0.0003757085487805	247.00
13	10/29/2008	0.0002786696621277	235.00

Figure 5

Return(date)

	A	B	C
1	**Date**	**Estimated Return**	**Return(date)**
2	09/15/2008	0.0002798024093023	-0.0471358951825517
3	10/06/2008	0.0001025773080189	-0.0385178716010297
4	10/10/2008	0.0001732615737705	-0.0117592755407068
5	10/16/2008	0.0002552908557377	0.0425074903066619
6	10/17/2008	0.0004193588322581	-0.0062128208108365
7	10/21/2008	0.0003545098779167	-0.0307996752587782
8	10/22/2008	0.0002512865661905	-0.0610125124339040
9	10/23/2008	0.0002895931437209	0.0126340908584046
10	10/24/2008	0.0001388618072519	-0.0345112376253978
11	10/27/2008	0.0003592420009615	-0.0317643167535386

Figure 6

Correlation Matrix

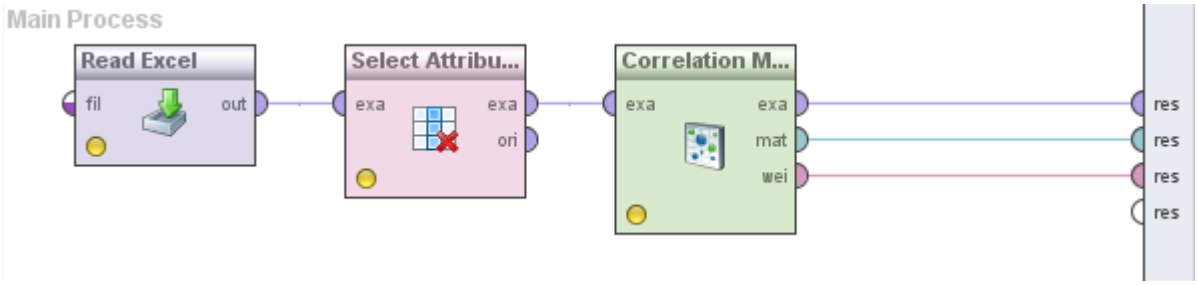

Figure 7

Correlation of Hypothetical Returns and Real Returns

Attributes	Estimated ...	Real Return
Estimated R	1	0.192
Real Return	0.192	1

Figure 8

Data loss from November 2009 to January 2010

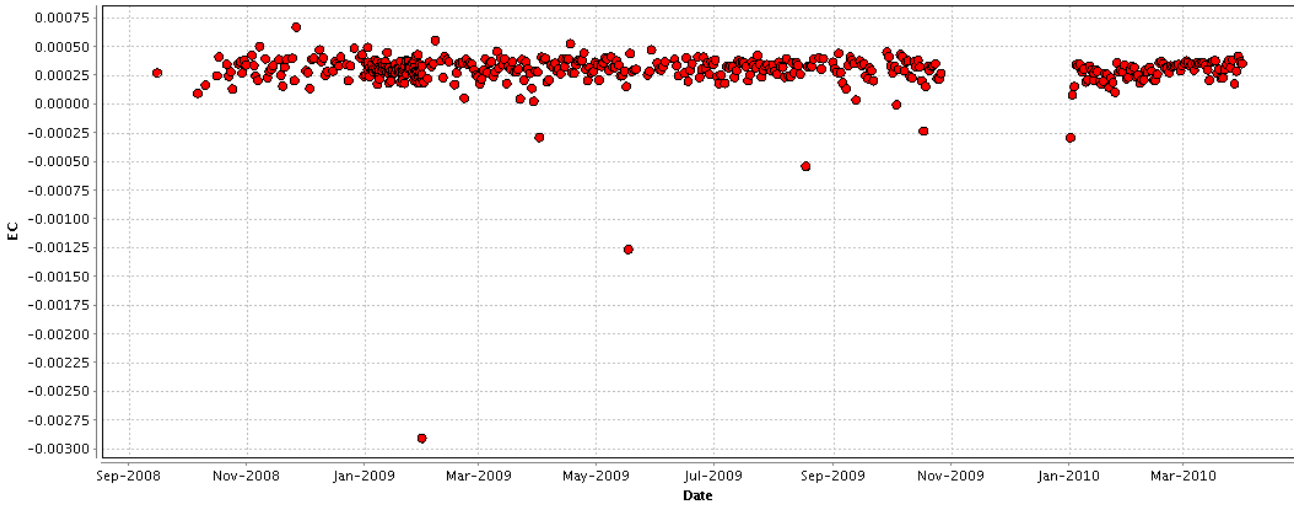

Figure 9

High fluctuation in the early periods of real return

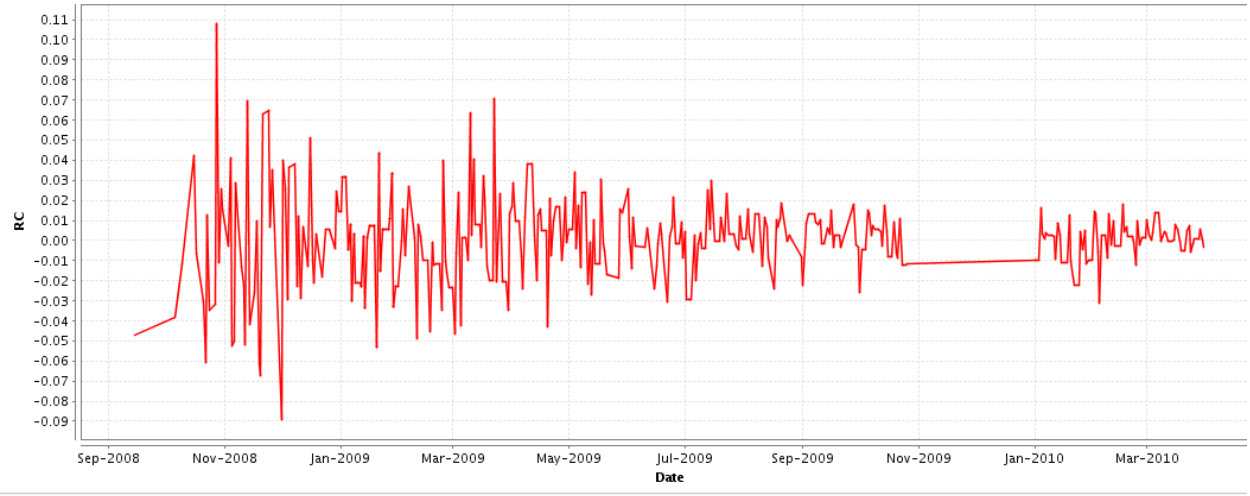

Figure 10

No visible high fluctuations in the early periods of estimated return

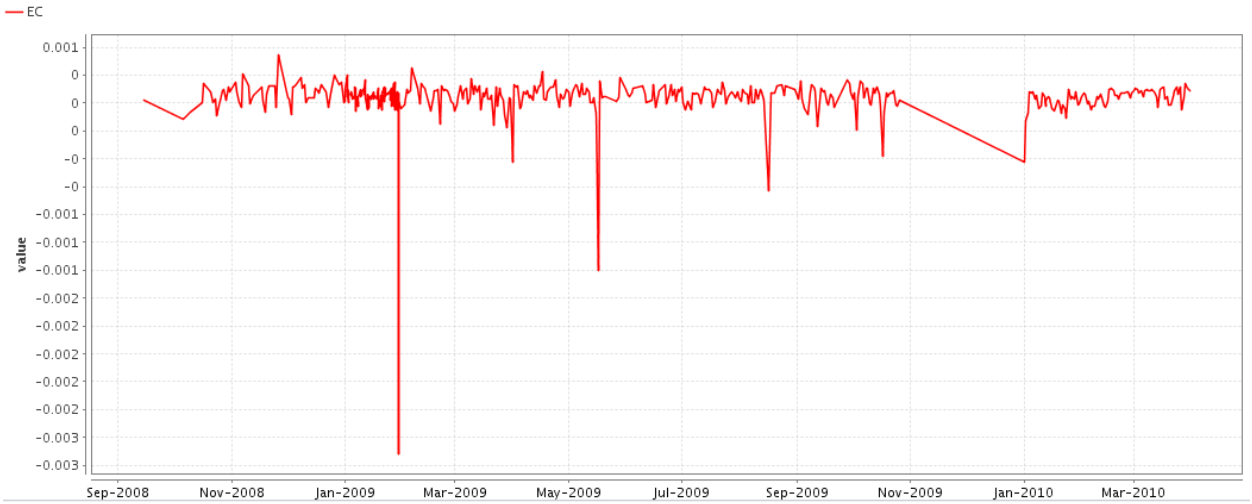

Figure 11

Words Association

2	acquiring	0.00319691
3	rally	0.00312329
4	acquired	0.00267631
5	breakout	0.00266934
6	acquires	0.00255162
7	lawsuit	0.00245513
8	merge	0.00232189
9	merger	0.00231289
10	higher	0.00229594
11	curve	0.00227158
12	climb	0.00215034
13	upside	0.00205971
14	squeeze	0.00185646
15	volume	0.00180592
16	beta	0.00168895
17	peak	0.00154205
18	announcement	0.00137009
19	past	0.00133629
20	strike	0.00133084
21	acquire	0.00131280
22	monthly	0.00129978
23	sold	0.00127177
24	high	0.00126894
25	chart	0.00121377
26	gold	0.00115793
27	gain	0.00114301

(a) Words associated with positive returns

	A	B
117	hate	-0.00008634
118	quarter	-0.00009115
119	loss	-0.00016146
120	debt	-0.00018902
121	euro	-0.00022953
122	china	-0.00023409
123	low	-0.00028172
124	bottom	-0.00037005
125	europe	-0.00040377
126	obama	-0.00040459
127	today	-0.00045693
128	vix	-0.00048354
129	announces	-0.00049764
130	bounce	-0.00052624
131	support	-0.00059819
132	usd	-0.00062090
133	announcing	-0.00064544
134	finance	-0.00069301
135	pound	-0.00070164
136	holdings	-0.00070749
137	down	-0.00080725
138	eps	-0.00086547
139	alpha	-0.00097745
140	futures	-0.00100139
141	year	-0.00120393
142	delta	-0.00145344
143	downside	-0.00165847

(b) Words associated with negative returns