PAPER • OPEN ACCESS

Sentiment analysis and relationship between social media and stock market: pantip.com and SET

To cite this article: P Padhanarath *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **620** 012094

View the article online for updates and enhancements.

Sentiment analysis and relationship between social media and stock market: pantip.com and SET

P Padhanarath[1], Y Aunhathaweesup[2] and S Kiattisin[3]
[1]Data Analyst, Sertis Co., Ltd., Bangkok, TH
[2]IT Management Lecturer, Faculty of Engineer, Mahidol University, TH
[3]IT Management Program Director, Faculty of Engineer, Mahidol University, TH

E-mail: praewmai.pah@student.mahidol.ac.th

Abstract. This research constructs a process flow of social media sentiment analysis and explains relationship between comments on social media and stock. One of the popular social media for Stock Exchange of Thailand investors is Pantip.com, a website providing service as a webboard with tagging feature. In this research, all posts tagged by 'Stock' on this website were crawled to files. Then comments were scraped and cleaned. For model training, some comments were labelled into three classes; positive, negative and neutral. Sentiment analysis model was constructed by Naive Bayes Classification technique. In evaluation, the model shown that it performed 74% accuracy. This model was utilised to classify comments into sentiments. When all comments were completely classed, sentiment types were counted by date. Finally, correlation matrices were constructed to find relationship between number of sentiments and stock. The research found that number of sentiments from social media relate to ADVANC and CPALL stock volumes. Moreover, the correlation always reaches to the peak on trading day then it gradually declines with the magnitude depending on the day length after trading day.

1. Introduction

In general, stock is known as a share owned by stockholders whose price is uncertain. Stock prices could move to any direction due to a change in demand and supply so many investors try to speculate from this variation. Indeed, one of factors that affect the amount of demand and supply is an expectation of the company in the future. All Investors expect to receive some profit in term of capital gain and/or dividend. Actually, the expectation is driven by confidence of investors. With a high level of confidence, investors tend to buy more stock. In contrast, investors may delay investing because of low confidence. Accordingly, confidence leads to action of investors and finally stock price tend to change due to these actions. Obviously, if most of investors want to buy more stocks, its price will go higher and vice versa. This shows that the level of confidence of investors have an directly impact to change of stock price.

In Thailand, Stock Exchange of Thailand (SET) has constructed ISI (Investor Sentiment Index) which is formed by sampling and survey base. ISI indicator represents overall sentiment or confidence of investors. Although the indicator can reflect what investors believe, the survey takes high cost and time consumption.

In the age of globalization, the whole world was connected by the internet. People seem to attend online social more and more. They are free to express their opinion and easily consume news as well. Topics on social media can be various from general publication to personal life, and stock is no exception. People also comment on stock and usually question about how stock today is. In return, some experts (their thought) not only reply back but also give suggestion and provide information as much as they know. Thus, social media is one of places that investors make discussion and casually leave their sentiments about stock where anyone can access directly via the internet. According to John R [1], his research found that when 2 or more people came across and started conversation about investing, they had exchanged thought, attitude, and mood together. As a result, it turned them on and made them think about their investment because communication was a psychological way heading to decision of people.

2. Related Work

2.1. What determines a stock price?

“The stock I am holding will its price go up or down?”, “Why the price change?”, “Could do I know or predict its change in advance?” these questions always come up to the investor’s mind who invests in stock. To answer those questions, the investor should understand what determine or affect to stock price is. And the thing help investor to clearly understand stock factors is ‘finding Intrinsic value of stock’ [2].

Intrinsic Value is derived from investors’ expectation on future return. So, the price they willing to pay now is worthy for things expected afterwards. Returns from expectation could be both ‘Dividend’ and ‘Capital Gain’. For example, if investors believe that investing in ‘stock A’ they can get high dividend, they will pay for it no matter how much it is. Due to increasing in demand, the price will go up. On the other hand, if investors consider that a price of stock A in the next 3 months will be lower, they will not buy it outright. They will wait until the price drops to a certain level which can cover the risk they may encounter. So the price of stock A goes down because of decreasing in demand. To summarize, the current market price is from 2 main factors; (1) Expected Return and (2) Risk that may occur. These factors affect to intrinsic value and market value respectively as shown on Figure 1.

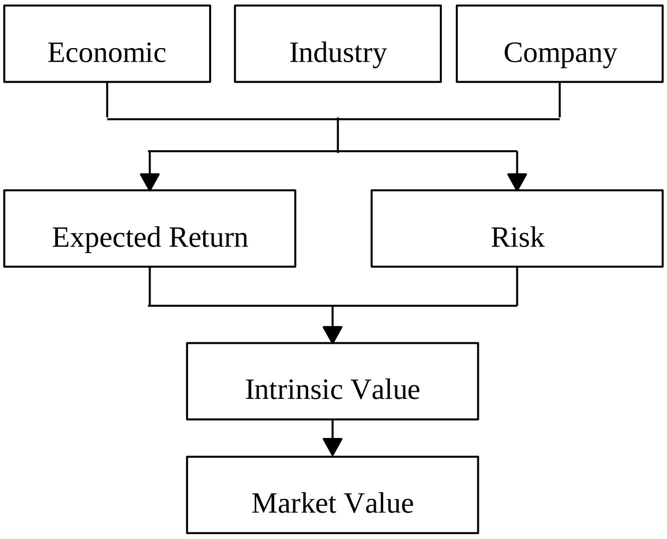

Figure 1. Determinant of Stock Price

The next question is “What determines dividend, selling price, and risk?”. The answer is tendency of future profit of company that performed by company’s operation, industry, trends of overall economic and stock market. These are factor analysed for determining stock price called ‘Fundamental Analysis’ [2]. The change in Stock Price Factor can be seen in the Figure 2.

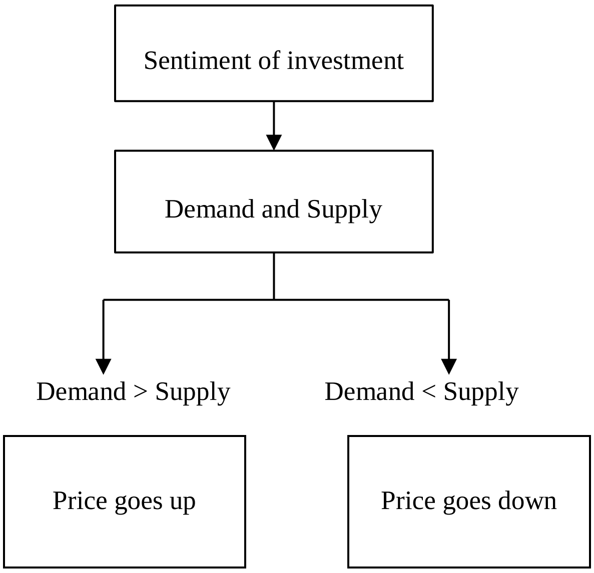

Figure 2. Change in Stock Price Factor

In addition, the thing that directly affects to change in stock price is demand and supply derived by sentiment of investors [3] [4]. Forecasting of investors leads to demand and supply of stock and then price would be changed by market mechanism.

As Peter Atwater’s belief [4], sentiment of investors has a huge impact of stock price changes, he said *“I believe that markets are not moved by corporate or economic data or even by external events but by us; by how we feel—our mood—and, importantly, by how changes in our mood drive our preferences and in turn the specific decisions that we make every day.”*

“At the risk of over-simplicity, I think of our own individual mood as our underlying confidence. It is how sure we feel, often unknowingly, about ourselves and the world around us.”

Thus, the important thing truly affects to change of stock price is investor’s confidence in investing. Besides, CMRI (Capital Market Research Institute) had been developing the tool to be a support information for investors which is called ISI (Investor Sentiment Index). They found that when value of ISI is higher than usual, it will affect to higher on SET index and also higher on other indicators i.e., volume, retail investor investment ratio. Moreover, small securities tend to be varied because retail investors usually make decision by their sentiment [5]. Thus, understanding sentiment could help investors to get more chance of success in stock investment.

2.2. Social Media: Pantip.com

Pantip.com or Pantip is a popular website providing webboard and chatroom services in Thailand. It was founded on October 7, 1996 and firstly its purpose is to create an online magazine but many users prefer using webboard for public comment and share. Therefore, Pantip has changed their business goal to be webboard and discussion forums. Thanks to the changing in business model, earning from advertisement in the website becomes the main income for Pantip.

Pantip has an impact to most of Thai people’s mindset and decision making since it is a big source of information used for searching and experience exchange in many topics. Until now, Pantip has the most traffic among Thai-language websites [6]. Recently, Pantip was also one of the top 5 websites in Thailand ranked by Alexa [7] on October 13, 2016 as shown in Table 1.

Table 1. Top 5 sites in Thailand ranked by Alexa

Rank	Site	Daily Time on Site	Daily Pageviews per visitor	% of Traffic From Search	Total Sites Linking in
1	Google.co.th	6:46	8.6	0.90%	7,837
2	Youtube.com	8:59	4.92	15.30%	2,532,990
3	Google.com	7:23	8.04	3.30%	3,356,487
4	Facebook.com	9:54	3.78	7.90%	6,685,204
5	Pantip.com	5:27	4.09	65.10%	4,956

Previously, Pantip had categorized forum types by chatting room. But now, tagging was added to one of their features. Tag can identify forum topic more precisely for example ‘Sinthorn room’, a room for financial discussion, can be separated by tag e.g., loan, investing, stock, etc. For ‘stock’ tags, this research found that there were more than 57,000 topics a year talking about stock on Pantip.com.

2.3. Sentiment Analysis

There is a vast pool of opinions and experiences that people had shared via the internet. This made it possible for one who wants to know what people think easier. These online opinions are valuable for a business whose performance mainly depends on customers’ preferences and that is where sentiment analysis initiated. As Theresa W said [8], Sentiment Analysis is a type of subjectivity analysis focused on identifying positive and negative opinions, emotions, and evaluations expressed in natural language. It has been a central component in applications ranging from recognizing inflammatory messages, to track sentiments over time in online discussion, to classify positive and negative reviews. For example, papers published in 2001 which authors were interested in analysing market sentiment, Das and Chen [9] and Tong [10], had used the term “sentiment” referred to automatic analysis of evaluative text and tracking of the predictive judgments [8]. In Thailand, the majority of sentiment analysis usages are product and service. One clear example is a paper in 2017 which authors focused on product review online by Support Vector Machine [11]. Another example is 2013 paper that written by Kanda [12]. She tracked online comments for developing new product and services that can satisfy customers’ demand.

3. Methodology

3.1. Proposed Model

This research propose model to be a workflow preparing sentiment analysis and correlation analysis as illustrated on Figure 3. To answer the question about relationship between stock and social sentiment, the research mainly focus on the whole process running to find the final correlation result. The scope of research is limited to top 10 stocks in SET with highest Market Capitalization (MC) because high MC stocks usually attract to investors more than low MC stocks significantly. After web crawling and scraping, texts are matched to their related stock and some matched texts are randomly selected to label.

On a single line, words used always have their hidden meaning so understanding of 'Stock Word' is needed for model training. After sentiment analysis, labelled texts are summarized to be daily sentiments, i.e. number of positive, negative, and neutral on each date. On the last step, correlation is performed to find the relationship on stock price, stock volume and sentiment texts.

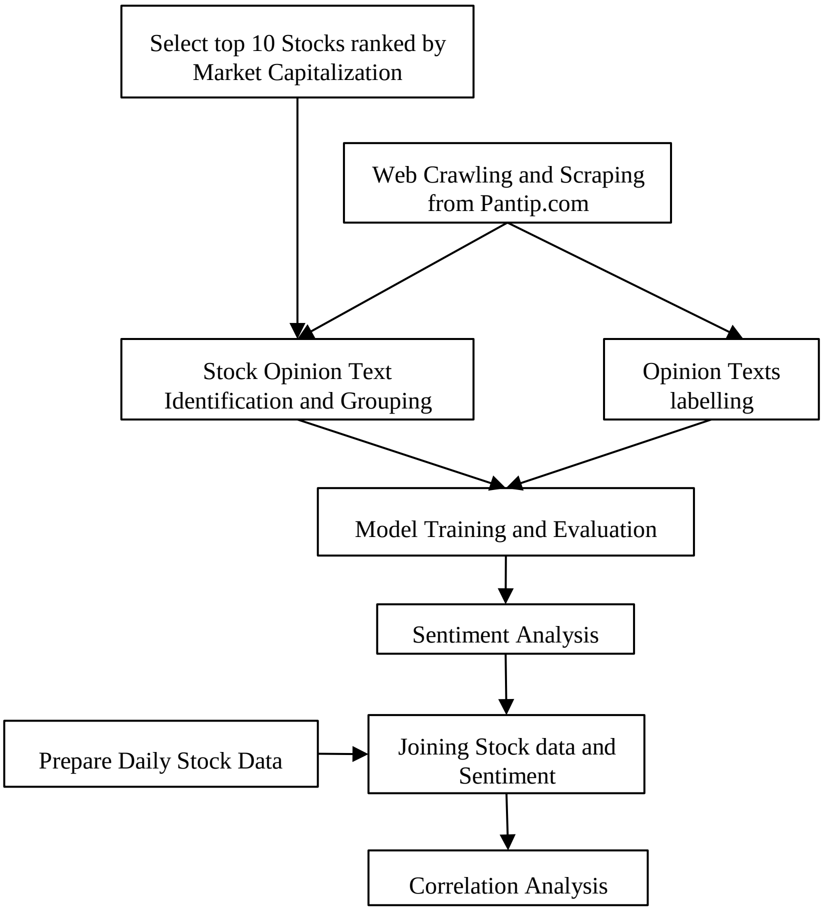
Select top 10 Stocks ranked by Market Capitalization
Web Crawling and Scraping from Pantip.com
Stock Opinion Text Identification and Grouping
Opinion Texts labelling
Model Training and Evaluation
Sentiment Analysis
Prepare Daily Stock Data
Joining Stock data and Sentiment
Correlation Analysis

Figure 3. Proposed Model

3.2. Sentiment Analysis: Naïve Bayes

Naïve Bayes is potentially good at serving as a document classification model due to its simplicity.[13] It was initiated fundamentally from Bayes' Rule, the statistical theory about utilizing probability to describe uncertainty in mathematical form.

If P(A)> 0, then the probability of event B occurs given that event A has occurred can be written as (1) [14]

$$P(B \mid A) = \frac{\mathrm{P(B \cap A)}}{\mathrm{P(A)}} \qquad (1)$$

P(A) = Probability of event A occurs

P(B) = Probability of event B occurs

P(B∩A)= Probability of both A and B have occurred

Refer to multiplication rule, the equation can be rewritten as (2)

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \qquad (2)$$

To classification, probability theorem can be used for solving Classification problems. It can predict and describe a result and also find the relationship between variables to construct conditional probabilities for each relation. Naive Bayes is an efficient method for classification i.e., Text Classification, it perform an uncomplicated approach where attributes are independent to each other.

$$P(A_1, A_2, \dots, A_n|C_j) = \prod_{i=1}^{n} P(A_i|C_j) \qquad (3)$$

i = 1, 2, 3,..., n and j = 1, 2, 3,..., n

Where C_j is a sentence j which contains words appeared in the bag of words in X = {A_1, A_2, ..., A_n}. Consequently, it can be written as P = (A_1, A_2, ..., A_n | C_j)

As a result, we got a Simple Bayes Classifier in equation (4) [11]

$$V_{NB} = argmax P(C_j) \prod_{i=1}^{n} P(A_i|C_j) \qquad (4)$$

Naive Bayes' requirement is a small amount of training data to estimate the parameters for learning the models. Also Naive Bayes is a descriptive and probabilistic machine learning, therefore, the results could be easily analysed and explained.[15] From S.L. Ting [13], the results show that Naïve Bayes is the best classifiers against several common classifiers (such as decision tree, neural network, and support vector machines) in term of accuracy and computational efficiency.

3.3. Model Training and Evaluation

Training data set needed for model training in this research is sentiment texts. Extracted texts from posts and comments in Pantip.com were scrapped and labelled. Finally, 3,000 Sentiment texts were labelled and used for model training (1,000 texts for each sentiment type; positive, negative, and neutral) can be seen in Table 2.

Table 2. Sample sentiment texts for model training

Text	**Sentiment**
Great. Stock is so strong	Positive
Tomorrow will be green	Positive
Ready to buy a stock	Negative
Prepare yourself, we are on the way down	Negative
Teach me to play the market please	Neutral
Will the market open tomorrow?	Neutral

For model evaluation, 10-fold cross validation is constructed where k=10 is seem to be a good compromise[16].This value of k is particularity attractive because it makes predictions using 90% of the data, making it more likely to be generalizable to the full data.

3.4. Correlation Analysis

3.4.1. Joining stock data and summarized sentiment data. Stock data, i.e., stock name, date, %price change, and volume, were matched with daily number of sentiments of that stock.

Table 3. Sample Stock and Sentiment Data by Date

Stock name	Date	%Price change	Volume (Shares)	Number of sentiments today		
				Positive	Negative	Neutral
ADVANC	01/07/2013	-0.49	2,880,000	13	28	12

3.4.2. Correlation Matrix Construction. The research studies the relation between stock and sentiment texts in the following scope shown in Table 4.

Table 4. Correlation Matrix Variables

Stock	Sentiment texts from Pantip.com
%Price Change Volume(Shares)	Number of positive, negative, and neutral texts 10 days before Number of positive, negative, and neutral texts 5 days before Number of positive, negative, and neutral texts 2 days before Number of positive, negative, and neutral texts today Number of positive, negative, and neutral texts 2 days after Number of positive, negative, and neutral texts 5 days after Number of positive, negative, and neutral texts 10 days after

3.4.3. Correlation Coefficient. The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables [17]. The correlation coefficient is scaled between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables. The Pearson correlation coefficient (r) is the most widely used and can be defined as follows. Suppose that there are two variables X and Y, each having n values $X_1, X_2,..., X_n$ and $Y_1, Y_2,..., Y_n$ respectively. Let the mean of X be $\bar{X}$ and the mean of Y be $\bar{Y}$. Thus the Pearson's r is

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

where the summation proceeds across all n possible values of X and Y [17].

4. Result

In this research, we select 10 stocks i.e., PPT, AOT, CPALL, ADVANC, SCC, PTTEP, KBANK, SCB, BDMS, and PTTGC, ordered by market capitalization respectively.

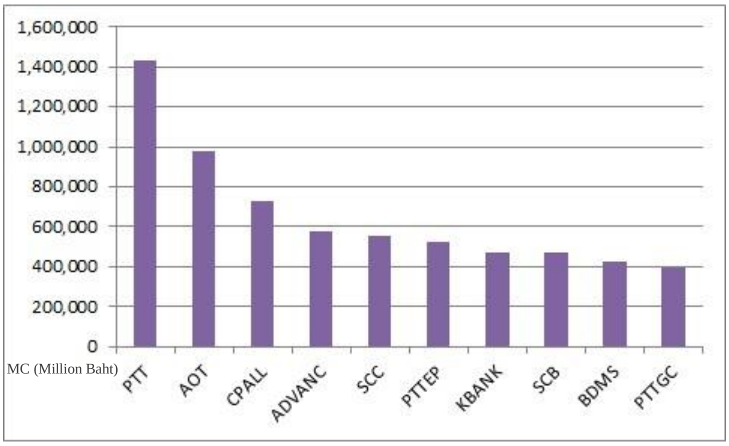

Figure 4. 10 Stocks ranked by market capitalization

As shown on Table 5 ADVANC got highest number of comments (107,638) followed by PTT (78,848) and CPALL (52,338) respectively.

Table 5. Number of comments by stock

STOCK NAME	NUMBER OF COMMENT
PTT	78,848
AOT	44,455
CPALL	52,338
ADVANC	107,638
SCC	27,341
PTTEP	50,815

KBANK	40,869
SCB	70,589
BDMS	12,359
PTTGC	34,880
TOTAL	**520,132**

4.1. Sentiment Analysis Model

In model evaluation, we got accuracy in average from 10-fold cross validation as shown on Table 6

Table 6. Average Accuracy of Sentiment Analysis Model

	Positive	Negative	Neutral
F1	0.686604	0.732277	0.731515
Precision	0.706469	0.810865	0.731515
Recall	0.668395	0.669217	0.815694
Accuracy	0.740000		

4.2. Correlation between sentiment on Pantip.com and stock.

The research found that there are 2 stocks tend to have relationship with sentiment texts, ADVANC and CPALL, the other 8 stocks are founded that their volume have a lower weak relationship on sentiment texts.

Table 7. Correlation Coefficient between ADVANC volume and sentiment

	5 days before			Today			5 days after		
	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral
Volume	0.37	0.42	0.4	0.64	0.69	0.62	0.37	0.42	0.4

For ADVANC, there is nearly a strong positive linear relationship between sentiment texts and volume on today. and lower moderate positive linear relationship on 5 days before and 5 days after.

Table 8. Correlation Coefficient between CPALL volume and sentiment

	5 days before			Today			5 days after		
	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral
Volume	0.49	0.5	0.47	0.79	0.77	0.73	0.52	0.48	0.47

For CPALL, there is a strong positive linear relationship between sentiment texts and volume on today. and moderate positive linear relationship on 5 days before and 5 days after.

5. Conclusion

Sentiment texts on Pantip.com relate to volume of some specific stocks; ADVANC and CPALL. Moreover, looking into correlation coefficient compared by number of days before and after the trading day, the result shows that correlation reaches to the peak on trading day then it gradually declines with the magnitude depending on the day length after trading day.

The result shows in the same way as the researches [3][4] that sentiment affects to demand and supply of stock. However, investor sentiment is not the only factor affecting the change in stock price, there are also other factors such as, economic outlook, natural disasters, politics, that can lead to movement of stock price. To work on, other social media and/or related factors can be included to make the model be more accuracy.

In general, data available is mostly unstructured and not organized in a predefined form. Most of this comes from text and sentiment analysis could allow companies to make sense of this unstructured text. Also, there are many areas using sentiment analysis e.g., social monitoring, voice of customer, product analytics, market research etc. Therefore, company can improve their process, product, and service to serve the customers' needs promptly.

References

[1] John R and Nofsinger 2005 *The Psychology of Investing* (New Jersey: Pearson Education, Inc) pp 140-141

[2] SET 2015 *Factors determine Stock Price* (Bangkok: SET) pp 1-3

[3] SET 2018 *Stock* Investor Classroom (Bangkok: SET)

[4] Peter A 2013 *Moods and Markets* (New Jersey: FT Press Upper Saddle River) p 12

[5] Sirijot J 2018 *Investor Sentiment Index: ISI (Electronic Material* Vol 3*)* pp 1-8

[6] Wikipedia Volunteer 2018 *Pantip.com* (Free Encyclopedia)

[7] Alexa 2017 *Top sites in Thailand* (Alexa Internet, Inc)

[8] Wilson T, Janyce W and Paul H 2009 Recognizing contextual polarity: Anexploration of features for phrase-level sentiment analysis *Comput. Linguist.* **35** pp 399-400

[9] Sanjiv D and Mike C 2004 *Yahoo! for Amazon: Extracting market sentiment from stock message boards* In Asia Pacific Finance Association Annual Conf. (APFA)

[10] Richard M and Tong 2001 *An operational system for detecting and tracking opinions in on-line discussion* In Proceedings of the Workshop on Operational Text Classification (OTC)

[11] Ravisuda T and Nives J 2017 Thai Sentiment Analysis of Product Review Online Using Support Vector Machine *Eng. J. Siam Univ.* **18** pp 2-10

[12] Kanda P and Pramote L 2013 Opinion Mining from Online Social Networks *Mod. Manage. J.* **11** pp 11-20

[13] S.L. Ting, W.H. Ip and Albert H.C. Tsang 2011 Is Naïve Bayes a Good Classifier for Document Classification? *Int. J. Soft. Eng. Appl.* **5** pp 37-46

[14] Samuel A and Broverman 2007 *ACTEX Study Manual SOA Exam P CAS Exam 1* (Winsted, Conn: Actex Publications) pp 57-58

[15] Choochart H, Alisa K, Pornpimon P and Kanokorn T 2013 *S-Sense: A sentiment analysis framework for social media sensing* (Nagoya: Asian Federation of Natural Language Processing) pp 6-13

[16] Payam R, Lei T and Huan L 2008 *Cross Validation* pp 1-6

[17] *Association between Variables* (Electronic Material) Chapter 11 pp 795-800