

## Sentimental Analysis on Big Data – On case of Financial Document Text Mining to Predict Sub-Index Trend

Johannes K. Chiang<sup>1, a</sup>, Chun-Cheng Chen<sup>2, b</sup>

<sup>1,2</sup>Department of Management Information System, National Chengchi University, Taipei City 11605, Taiwan

<sup>a</sup>email: jkchiang@nccu.edu.tw, <sup>b</sup>email: 104356507@nccu.edu.tw

**Keywords:** Sentimental analysis, Big Data, LDA, SVM, Taiwan Electronic Sub-Index Trend

**Abstract.** This research analyzed the potential emotion by sentimental analysis in large volume of financial text documents about Taiwan electronic industry to predict the stock trend. In recent researches about sentimental analysis, supervised method was proven to be able to reach high accuracy, but the training set of supervised method should be classified by manpower and couldn't discover the unknown category. So this research put forward a solution which mixed supervised and unsupervised methods. First, we introduce unsupervised method to find out the topics of documents. Then we calculated the sentimental index to judge the document's emotional direction. After that, we find out which theme documents' sentiment index are leading indicators in Taiwan electronic sub-index (TE). Finally, we used supervised method by integrating the sentimental index of leading indicators with other 24 indirect sentimental indexes to build the prediction model of TE. By result, we found that LDA model has better cluster performance than TFIDF-Kmeans model, and also has higher accuracy than NPMI-Concor model on classification. By comparing sentimental index with MACD, we proved that the trend of sentimental index and TE to each other is more similar than MACD line and TE to each other. We also discovered that the sentiment indexes from enterprise operation and macro-economics topics are leading indicators and found that the prediction model of TE which includes the sentiment index is better than which only includes the technical indicators.

### Introduction

Sentimental analysis, also called Opinion Mining, is meant to find out the author's thought, opinion and attitude about events or topics by computer analysis [1]. Applying sentimental analysis on text can find out public's potential emotion and know their opinions about events fast without wasting lots of manpower [2]. Many sentimental analysis researches use the CKIP, MMSEG or Jieba tokenization system to process the text like removing the stop words, part-of-speech tag and text chunking. Then calculate the sentimental score by comparing the words with the sentimental dictionary [3]. The numbers of words in dictionary affect the performance of word's polarity judging and sentimental score calculating [4]. Some nouns or words have different polarity in different fields. For example, "put" and "call" are verbs without polarity in many fields, but they have polarity in the financial field. This situation influences the classification performance of sentiment dictionary. Researchers find the proper nouns by calculating the co-occurrence between classified documents features and sentimental words [5].

In 1990, the researchers of behavioral finance found that public emotion could affect the stock price to deviate its normal price [6]. Many researchers tried to react to investor's emotion by indirect sentimental index like overbought, oversold, and turnover rate of market and proved that investors' emotion can affect the stock price. American Investor Trust used institutional investors' professional suggestions to build the Investors' Intelligence Sentiment Poll of New Rochelle. They found it has positive correlations with DJIA, S&P500, and NASDAQ [7]. Bollen *et al* [8] classified 100 million posts on Twitter to six kinds of emotion such as calmness, circumspection, belief, activity, kindness and happiness to predict the DJIA every day. The result shows that DJIA will raise if there are more calmness documents in that day; otherwise, DJIA will fall if there are less calmness documents in that day. Its accuracy is 87.6%.

Liu et al [9] used TFIDF and K-means to cluster the Sogou website's data then had a good topic classification. Yei [10] analyzed a listed semi-conductor companies' financial report from 2002 to 2011 by TFIDF and K-means to find out the company performance. Lin [11] used Concor to build the topic model, and used NPMI and Concor to judge the topic of the comments about various messenger Apps. Its accuracy is 96%. Blei, Ng and Jordan [12] put forward to the LDA model. LDA is an unsupervised learning model, and its main idea is that documents exhibit multiple topics. Each article is composed of many topics and each topic is composed of many words. Griffiths et al [13] used the topic model's perplexity to decide the number of topics. Perplexity is used to evaluate the probability model's performance. The perplexity is lower, the performance is better. Hung [14] used LDA on the large numbers of e-book in the library to find out the reader's preference. Chang [15] used LDA on the Plurk posts to find out public's preference and popular topics.

### The Approach of Sentimental Analysis

Our research process is separated into 6 parts, including data collection, data processing, topic tagging, sentiment orientation tagging, visual analysis, and building classification model (Fig1).

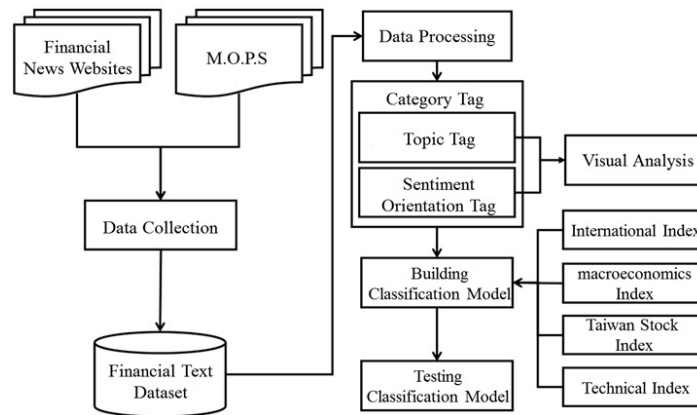


Fig.1. Research Process

In data collection, we collected financial text documents from cnYES, CNA News, China Times, and M.O.P.S. in 2014.

Data processing encompasses Tokenization, Part-of-Speech Tag, Negation Processing, Part-of-Speech Filtering, and Calculating Terms Frequency. We used Jieba to tokenize documents to words, and attach part-of-speech tag to each term. Negations shown before or after the predicates would cause the contents to have opposite meaning. We search these 2 positions to see negations if exist or not. To make the topic tag and sentiment orientation tag more precisely, we only keep nominals (common nouns/Na) for topic tag and predicates (VH-VL) for sentiment orientation tag, and remove others like stop words. In order to distinguish the word's importance, we would calculate the document frequency, term frequency and inverse document frequency. Using them to quantize the text documents and build the vector space model in cluster or classification steps.

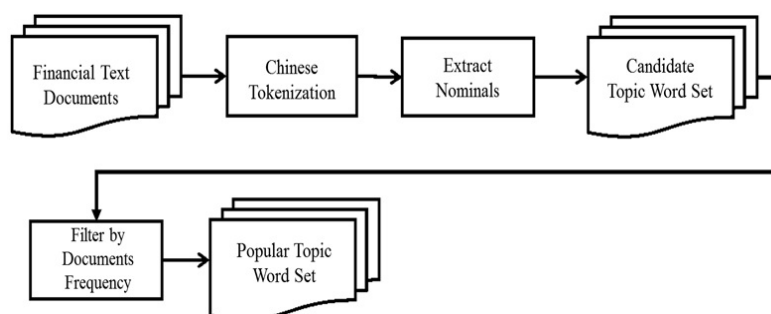


Fig.2. Finding Popular Topic Words

While nominals show the topic of contents, we found popular topic words (Fig.2) by extracting them from the contents and added to candidate topic word set. The topic is more important, and the frequency of discussion is higher. So we removed the words which documents frequency is lower than 20 [4].

Then, we used clustering methods, viz. TFIDF-Kmeans, NPMI-Concor and LDA, to build three kinds of topic models and compare their performance. We found that TFIDF matrix of model is too scattered to cluster, because of large difference between the numbers of documents and topic words. Thus, we tried to use the NPMI-Concor and LDA to do topic extraction and compare with TFIDF-Kmeans. Lastly we found out the best topic model and used it to tag the documents one or few topics. If the document does not exist any topic words then remove it.

In Sentiment Orientation Tag phase, we at first expanded the word set, and classified the documents to positive or negative. Then, we found out the high frequency words from positive and negative documents separately, and expanded these words to the word set after artificial filtering. Second, we calculated the Sentiment Index, and compared all predicates w's polarity of each document  $d_i$  with sentimental word set. The word w gets a 1 point if it shows in the positive word set. On the other hand, it gets a -1 point if it shows in the negative word set (Eq. 1). Then, checking w if exists the negative tag or not. If it has negation, it gets the -1 weight (Eq. 2). Then, we summed the predicates w's point and got the document  $d_i$ 's sentiment orientation score  $SO\ Score(d_i)$  (Eq. 3). We used Z-score to normalize the  $SO\ Score(d_i)$  and got the sentiment index (Eq. 4). Finally, according to  $Sentiment\ Index(d_i)$ , we tag the document a sentiment orientation as positive, negative and neutral (Eq. 5).

$$Orientation(w) = \begin{cases} 1 & \text{if positive} \\ -1 & \text{if negative} \end{cases} \quad (1)$$

$$Negation = \begin{cases} 1 & \text{if negation not exist} \\ -1 & \text{if negation is exist} \end{cases} \quad (2)$$

$$SO\ Score(d_i) = \sum_{w \in d_i} Negation \times Orientation(w) \quad (3)$$

$$Sentiment\ Index(d_i) = \frac{SO\ Score(d_i) - \mu}{\sigma} \quad (4)$$

$$Sentiment\ Index(d_i) \begin{cases} > 0 \rightarrow \text{Positive Orientation} \\ = 0 \rightarrow \text{Neutral} \\ < 0 \rightarrow \text{Negative Orientation} \end{cases} \quad (5)$$

This research used visual analysis to prove that the sentiment index could predict the trend of TE. We compared the sentiment index with MACD, a technical index most investors used, and proved that the sentiment index is more accurate to predict the trend of TE than MACD does. Finally, we integrated the leading indicators with other 24 indirect sentimental indexes to build the vector space model. Then we used kNN, Naïve Bayes, SVM, and logistic to build the classification model and tested performance by using precision, recall and f-measure.

## Test Results

The topic model clustering result of TFIDF-Kmeans is not well (Table 1). There are 95% words clustered to group 3. The inhomogeneous distribution of words would affect the result of judging topic. In NPMI-Concor topic model (Table 1), the words distribution of NPMI-Concor is more homogeneous than TFIDF-Kmeans. However, the words in each group include so many topics that it is hard to generalize one topic to represent these words and affect the performance of topic tagging. In LDA topic model (Table 1), the words distribution of LDA is more homogeneous than

TFIDF-Kmeans too, and its words in each group have high correlation. It is easy to generalize to one topic for each group. We summarized these groups to four topics: stock market, enterprise operation, industry, and macro-economic.

Table 1. Topic model clustering result compared by TFIDF-Kmeans, NPMI-Concor, LDA

Cluster method	Group1	Group 2	Group 3	Group 4
TFIDF-Kmeans	2%	2%	95%	1%
NPMI-Concor	26.4%	27.2%	22.4%	24%
LDA	20%	35%	26.4%	18.4%

We compared the performance of NPMI-Concor and LDA on classification (Table 2). LDA topic model is more homogeneous than TFIDF-Kmeans in words distribution, and has a 5% higher accuracy than NPMI-Concor on classification. So, we chose LDA topic model to do the topic tag.

Table 2. The classification performance of NPMI-Concor and LDA

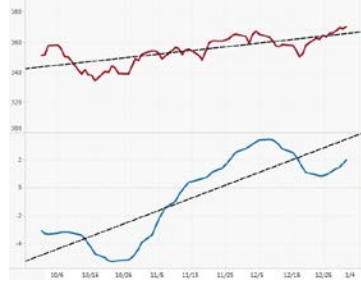

NPMI-Concor Topic Model				LDA Topic Model			
Group	Precision	Recall	F-measure	Topic	Precision	Recall	F-measure
Group 1	96.1%	97.9%	97%	Stock Market Topic	99.4%	98.6%	99%
Group 2	97.6%	91.9%	94.6%	Enterprise Operation Topic	97.9%	99.3%	98.6%
Group 3	99.5%	88.2%	93.7%	Industry Topic	98.8%	97.5%	97.9%
Group 4	90.2%	97.8%	93.9%	Macroeconomic Topic	99.8%	97.4%	98.6%

For Sentiment Orientation, we computed the sentiment index by comparing with sentiment word set, and tagged its sentiment orientation (Table 3). We found that there are only 12% documents cannot be judged its sentiment orientation. We inferred that some numerical documents include rare predicates to judge its sentiment orientation. And some documents exist few predicates to judge its sentiment orientation, but some nominals have polarity too.

Table 3. The Result of Sentiment Orientation Tag

	Positive	Negative	Neutral
Paper	7914	14091	3251
Percentage	31%	57%	12%

Table 4. Trend Line Analysis of Sentiment Index and MACD Result

	Trend Line Analysis of TE and MACD	Trend Line Analysis of TE and Sentiment Index
The data collected from Oct. to Dec. in 2014. (Red line is TE, blue line is MACD, and green line is sentiment index.)	 <p>cosθ:0.9717 angle:13°</p>	 <p>cosθ:0.9893 angle:8°</p>

To prove that using sentiment index could predict TE accurately, we drew the trend line and compared with MACD. We found that the angle between sentimental index trend line and TE trend line is less 5° than the angle between MACD trend line and TE trend line (Table 4).

Although sentiment index can react to the trend of TE accurately, only leading indicators could be used for predicting. We integrated sentiment index with topic tag, and drew line graph of sentiment index and TE. With parallel transport and correlation coefficient, we found out which topic's sentiment index is leading indicator (Table 5).

We found that only enterprise operation topic and macro-economic topic documents are leading indicators. They will react to the trend 4 to 6 days earlier, and its correlation coefficient is 0.87 to 0.9. Hence, we used enterprise operation topic and macro-economic topic's sentiment index to build the classification model, and built classification model through kNN, Naïve Bayes, Support Vector Machine (SVM) and Logistic algorithm. Then compared the performance of the model includes sentiment index with the model only includes technical indicators (RSI and KD) (Table 6).

Table 5. Line graph of sentiment index and TE (Orange line is TE and blue line is sentiment index)

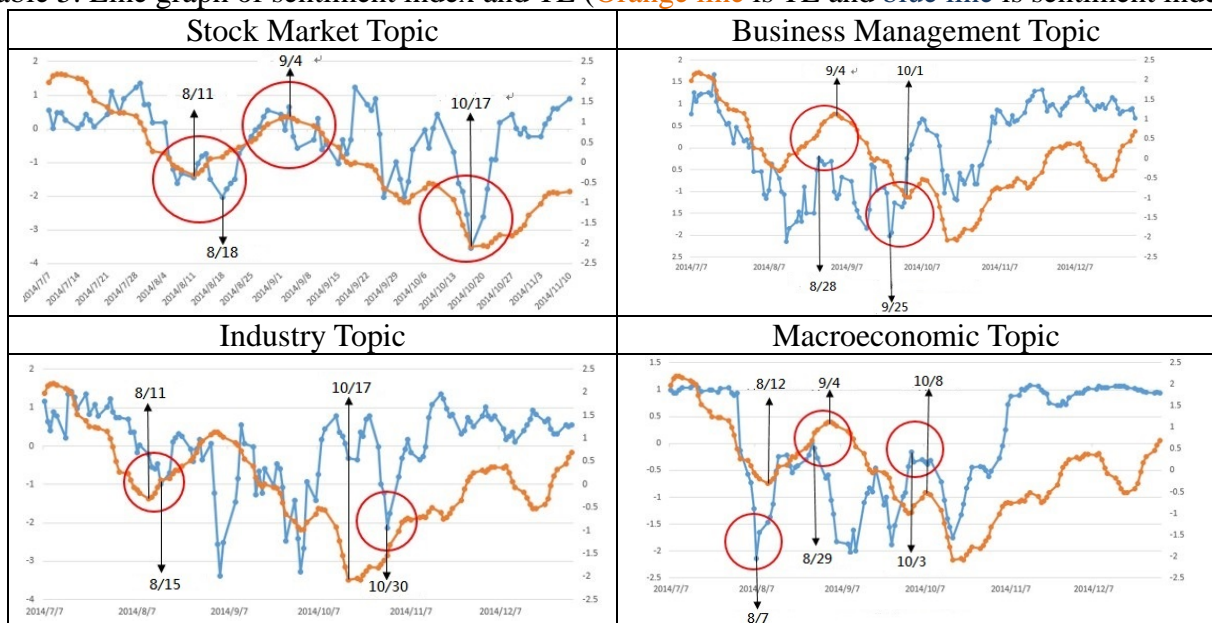


Table 6. Classification Model Result

Model only include technical indicators				Model include sentiment index		
kNN						
Class	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Avg.	43.6%	45.3%	43.5%	58.8%	56.3%	55%
Naïve Bayes						
Class	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Avg.	47.6%	49.5%	46.3%	58.4%	58.4%	58.4%
SVM						
Class	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Avg.	57.4%	55.7%	55.3%	62%	61.9%	61.4%
Logistic						
Class	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Avg.	44.9%	44.3%	44.2%	53.8%	52.3%	52.5%

Table 7. Performance of Classification model include indirect sentiment indicators

Classification model include indirect sentiment indicators			
Class	Precision	Recall	F-Measure
Avg.	71.3%	70.6%	70.7%

We found that SVM model has the best performance, thus used SVM to build the classification model. To enhance the accuracy, we added 24 indirect sentiment indicators according to Tsai's [16]

research result. After optimizing, the accuracy is up to 71%. (Table. 7)

## Conclusion

In sentiment orientation tagging, this research used artificial way to filter few high frequency words, and we can prevent to expand the word set from artificial interference by trying in other ways. We found that only few nominals have polarity, and we can try to add them to sentiment word set to make the result more accurate. And, this research used the parallel transport and Pearson product-moment correlation coefficient to find out which topic's sentiment index is leading indicator. In the future, we could try to use Cross Correlation Function (CCF) or Autocorrelation Function (ACF), and use other field's data to predict the trend of different financial product.

## References

- [1] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, 2008.
- [2] L.C. Chang. A Study of the Relevance between News Topics & Public Opinion and Stock Prices in Big Data, 2014.
- [3] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede. Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, 2011.
- [4] Y.L. Lin. Research into App user opinions with Sentimental Analysis on the Google Play market, 2014.
- [5] Du Jiazhong, Xu Jian, Liu Ying. Research on Construction of Feature-Sentiment Ontology and Sentiment Analysis, 2014.
- [6] J. Bradford De Long, Andrei Shleifer, Lawrence H. Summers, Robert J. Waldmann. Noise Trader Risk in Financial Markets, Journal of Political Economy, 1990.
- [7] Wayne Y. Lee, Christine X. Jiang, Daniel C. Indro. Stock market volatility, excess returns, and the role of investor sentiment, Journal of Banking & Finance, 2002.
- [8] Johan Bollen, Huina Mao, Xiao-Jun Zeng. Twitter mood predicts the stock market, 2010.
- [9] LIU Peng, TENG Jia-Yu, ZHANG Guo-Peng, HU Yan-Jun, HUANG Yi-Hua. Study of parallelized k-means algorithm on massive text based on Spark, 2014.
- [10] Y. H. Yei. Using text-mining analysis on qualitative information to predict companies' financial performance, 2012.
- [11] Y. L. Lin. Research into App user opinions with Sentimental Analysis on the Google Play market, 2014.
- [12] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation, Journal of Machine Learning Research, 2003.
- [13] Thomas L. Griffiths, Mark Steyvers. Finding scientific topics, Proceedings of the National Academy of Sciences, 2004.
- [14] C. Y. Hung. An Approach to eBook Topics Trend Discovery Based on LDA and Usage Log, 2012.
- [15] Jih-Wei Chang. Apply LDA with Topic Categories on Plurk and User Sentiment Analysis, 2014.
- [16] C.T Tsai. The Study of the Forecasting of the Stock Prices and Trend for the Electronic Industry in Taiwan, 2007.