Summer 2017

Forecasting Company Revenue Trend Using Financial News

Hsin-Ching Huang
National Sun Yat-sen University, volerhaut15@gmail.com

San-Yih Hwang
National Sun Yat-sen University, syhwang@mis.nsysu.edu.tw

Shanlin Chang
National Sun Yat-sen University, shanlin.c.fish@gmail.com

Yihuang Kang
National Sun Yat-sen University, ykang@mis.nsysu.edu.tw

Follow this and additional works at: http://aisel.aisnet.org/pacis2017

Recommended Citation
http://aisel.aisnet.org/pacis2017/193
Forecasting Company Revenue Trend Using Financial News

*Indicate Submission Type: Completed Research Paper*

**Hsin-Ching Huang**  
Department of Information Management, National Sun Yat-sen University  
Kaohsiung, Taiwan  
volerhaut15@gmail.com

**San-Yih Hwang**  
Department of Information Management, National Sun Yat-sen University  
Kaohsiung, Taiwan  
syhwang@mis.nsysu.edu.tw

**Shanlin Chang**  
Department of Information Management, National Sun Yat-sen University  
Kaohsiung, Taiwan  
shanlin.c.fish@gmail.com

**Yihuang Kang**  
Department of Information Management, National Sun Yat-sen University  
Kaohsiung, Taiwan  
ykang@mis.nsysu.edu.tw

**Abstract**

Text mining has emerged as an important suite of techniques in recent years and found its way into many applications. In the finance area, most recent text mining-based studies focus on the prediction of the stock market trend or the detection of company bankruptcy/fraud. Other important economic indicators of the companies, such as revenues, are seldom addressed. Yet these indicators could be quite important and reflect the financial status of the company’s cash flow and market share. In this paper, we adopt a lexicon-based approach that first builds several lexicons of different types, including sources, entities, aspects, sentiments, and past times. Twelve sentiment features are identified as predictors of revenue trend and a lexicon-based method for determining the sentiment of each feature is proposed. In addition, one more feature computed using ARIMA based on previous revenue data is incorporated. Our experimental results using news articles of the seven Taiwan-based, major PC manufacturing companies demonstrate that both financial news articles and previous revenue data are important for accurately predicting revenue trend. The prediction model constructed using the proposed approach is able to predict revenue trend with accuracy of more than 80%.

**Keywords:** financial news, sentiment analysis, revenue prediction, time series, ARIMA model
Introduction

Recent researches on analyzing the huge volumes of unstructured data have received increasing attention, and the developed text-mining techniques have been applied in many areas, such as biomedical (Hogenboom et al. 2016), marketing (Büschen and Allenby 2016), and finance (Wang et al. 2016). In finance domain, for example, financial statements, annual reports, and news are important sources for operators, managers, and investments to understand company’s recent performance and grasp market trends. Nevertheless, the ever increasing large amounts of unstructured data become more and more difficult and time-consuming for human comprehension, which calls for text mining techniques for (semi-) automatic processing. Some studies focus on the financial status of companies and develop techniques for detecting the possibility of bankruptcy or fraud (Glancy and Yadav 2011; Jans et al. 2010; Ravisankar et al. 2011; Zhou and Kapoor 2011). Some other studies employ sentiment analysis on financial news to anticipate company stock trends (Boudoukh et al. 2013; Schumaker et al. 2012; Wang et al. 2016). Xu (2014) explores multiple data sources by combining the conventional time series ARIMA model analysis technique with information from the Google trend website and news from the Yahoo finance website to predict weekly changes in stock prices. That is to say, the main research focuses on financial text mining are about bankruptcy/fraud detection and stock price trend prediction.

Revenue is an important factor in determining the success of a company. Kim (2001) demonstrated that profitability and revenue are the most common factor of organization goals in the context of economic value. Rust et al. (2002) found that firms adopting primarily a revenue expansion emphasis perform better than those who emphasize on cost reduction. In Ma et al. (2006), a network-based approach is developed to discover company revenue relations (CRRs) from financial news. A CRR is simply a binary value that indicates which company in the associated pair has higher revenue. In Our previous work (Hsieh et al., 2016) proposes an approach that predicts a company’s revenue trend by analyzing financial news is proposed. However, the proposed approach is still primitive and very limited evaluation results are reported. Besides, the proposed prediction model is purely based on sentiments expressed in recent financial news and does not consider the time series revenue data. Our work extends the previous work by considering more textual features and incorporate information from previous revenues by applying time series method for constructing a new prediction model. Our experimental results using news and revenue data (in traditional Chinese) about the seven major Taiwan PC manufacturing companies demonstrate that our proposed approach more accurately predicts their quarterly revenue trends with precision and recall values higher than 80%.

This paper is organized as follows. The next section will present the process of our proposed approach, which is a non-supervised, lexicon-based approach. We then describe our lexicon building process and how to derive entities, aspects, and sentiment from news content. The prediction model construction method will be subsequently presented. We collect news and financial report data about seven major PC manufacture companies in Taiwan and evaluate our approach using the collected data. The evaluation results will be reported. Finally, we summarize this paper and give future research directions.

The Skeleton of Our Approach

The skeleton of our approach is shown in Figure 1. We first decide on the data sources and collect the data. Then we define the schema of news and events that will be used to house the processed data. In the next step, we construct lexicons for various aspects and sentiment, and subsequently, the sentiments pertaining to various aspects will be derived. Such feature sentiments and previous time-series revenue data will be regarded as features for building the prediction model for forecasting revenue trend of a target company. The details of each step will be described in the following sections.

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Schema Design</th>
<th>Lexicon Construction</th>
<th>Sentiment Analysis</th>
<th>Prediction Model Construction</th>
</tr>
</thead>
</table>

Figure 1. Skeleton of Our Approach
For a target company, we consider news articles that describe entities relevant to the target company, e.g., influential subsidiaries and influential joint venture, as well as the industry to which the target company belongs. These news articles will be collected. In addition, the revenue data that are revealed in the quarterly financial reports of the target company will also be collected.

**Schema Design**

Text mining techniques aim to convert unstructured text data into some structured formats. We thus design the schema for hosting the structured data after processing news articles. Our news message schema consists of the following attributes:

- **Source**: The financial service that announces a certain message. Example financial services include Morgan Stanley, Barclays Stockbrokers, and Credit Suisse Group.
- **Date**: The date of the announcement by the source, using the format of ‘yyyy-mm-dd’. If no date is explicitly mentioned in a news, we will simply use the news date as its value.
- **Entity**: An entity is intended to a subject which could be an industry, a company, or a subsidiary mentioned in a message.
- **Aspect**: An aspect describes a particular perspective about the entity. For example, for a company entity, there are various aspects such as product evaluation, shipment, order, earning, and company evaluation. This field records the aspect a message is about.
- **Sentiment**: The field records the sentiment value of a message, which could be +1 (positive) or -1 (negative).
- **Agency**: It records the agency of the corresponding news article, e.g. China Times, Liberty Times, United Daily News, Central News Agency, and Apple Daily News.
- **NewsDate**: The published date of the corresponding news article, using the format of ‘yyyy-mm-dd’.

For example, for the following news article we crawled from Apple Daily News, after applying the method for identifying aspects and their sentiments, which will be described later, we can convert it into news message record as shown in Table1. It describes a financial observation released by Morgan Stanley (摩根士丹利) about the positive shipment in Laptop industry:

摩根士丹利證券指出，本季NB出貨約3440萬台，季增9%，年增1%。蘋果日報. 2014-07-14.

<table>
<thead>
<tr>
<th>Source</th>
<th>Date</th>
<th>Entity</th>
<th>Aspect</th>
<th>Sentiment</th>
<th>Agency</th>
<th>NewsDate</th>
</tr>
</thead>
<tbody>
<tr>
<td>摩根士丹利 (Morgan Stanley)</td>
<td>2014-07-14</td>
<td>筆記型電腦產業(Laptop industry)</td>
<td>出貨 (Shipment)</td>
<td>+1</td>
<td>蘋果日報 (Apple Daily)</td>
<td>2014-07-14</td>
</tr>
</tbody>
</table>

We define an event as an announcement made by a specific source on a particular date about some aspect of an entity. An event can actually be reported by several news messages, with the common entity, aspect, data, and source. Thus, we can obtain the information about events by grouping them based on **Source**, **Date**, **Entity**, and **Aspect** from the news message table. Next, we will define event schema as **Source**, **Date**, **Entity**, **Aspect**, **Sentiment**, and **Count**. The meaning of each field is described as follows:

- **Source**, **Date**, **Entity**, **Aspect**: the same definitions as shown in news message schema.
- **Sentiment**: The aggregated sentiment.
- **Count**: The number of messages for the event.
Note that the first four attributes serve as the dimension attributes, and the last two attributes, namely sentiment, and count are measurement attributes. Table 2 shows the event record obtained from the previous example news message, assuming that there are three news articles that report the same observation:

<table>
<thead>
<tr>
<th>Source</th>
<th>Date</th>
<th>Entity</th>
<th>Aspect</th>
<th>Sentiment</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>摩根士丹利</td>
<td>2014-07-14</td>
<td>筆記型電體回業</td>
<td>出貨</td>
<td>+1</td>
<td>3</td>
</tr>
<tr>
<td>(Morgan Stanley)</td>
<td></td>
<td>(Laptop industry)</td>
<td>(Delivery)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Lexicon Construction**

To determine values of the attributes Entity, Aspect, Source, Time, and Sentiment for news messages, we propose to adopt a lexicon-based approach. A lexicon is constructed for each of these attributes.

**Entity Lexicon**

As mentioned, we consider news articles that involve the target company as well as the relevant industry and the related companies. Thus, the entity lexicon contains terms in three categories, namely industry, company, and related companies. As entity terms in each category vary across different target companies, we manually construct the entity lexicon for a given target company. Table 3 shows some sample entity terms using “和碩” (Pegatron) as the target company:

<table>
<thead>
<tr>
<th>Entity Category</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>筆電產業、下游硬體製造產業、PC 產業、科技產業、手機代工產業、面板產業</td>
</tr>
<tr>
<td>Company</td>
<td>和碩、和碩聯合、4938</td>
</tr>
<tr>
<td>Related Companies</td>
<td>和碩子公司、華擎、華擎科技、景碩、晶碩、鎧勝、永碩</td>
</tr>
</tbody>
</table>

**Aspect Lexicon**

Some news articles directly report the predicted revenue and/or profit of a company and they are classified in the aspect earning (收益). Others may mention about the order (訂單) or shipment (出貨) of a company which obviously impact the revenue. In addition, several studies have confirmed the relationship between reputation (評價) and revenue (Macias et al. 2008). Thus, we collect aspect terms that are related to revenues in four categories, namely order, shipment, earning, and reputation. Table 4 shows some sample terms in our aspect lexicon:

<table>
<thead>
<tr>
<th>Aspect Category</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order</td>
<td>訂單、接單、下單、接獲</td>
</tr>
</tbody>
</table>
**Source Lexicon**

We look into the news articles in our corpus and identify some simple rules to extract the news sources mentioned in the news. For example, words such as指出(indicate), 表示(show), and 看好(be optimistic) usually follow some source organization names, so we apply these rules and obtain a set of source terms. Note that source terms with the same prefix will be grouped into a single group, e.g. 群益投顧 and 群益證券 are regarded as 群益. As a result, we obtain totally 24 sources, including 里昂, 瑞信, 法說會, 花旗, 法人, and 外資.

**Past-time Lexicon**

The method proposed by Hsieh (2016) does not consider time aspect and may mistakenly derive the sentiment for news articles that describe the performance of a company in previous seasons. For example, if the information retrieved from the news is about the company's performance of last year (去年), it is not expected to impact the revenue of the company in the coming season. Thus, we construct a past-time lexicon and set the distance threshold as 10. For each sentence, if the previous 10 words and the next 10 words of the entity involve any terms in this past-time lexicon, this sentence will be regard as irrelevant and discarded. There are totally 23 terms in the past-time lexicon, including 去年(last year), 上半年(first half of the year), and 前一季(previous quarter). For example, consider the following example news article we crawled from United Daily News.

和碩昨晚(10)日公布去年合併營收7,554億元，比前年大增56%，優於公司預期。聯合新聞網。2013-01-11

Such a sentence will be discarded because it's about last year's (去年) performance, even though its entity (和碩) is indeed our target company.

**Sentiment Lexicon**

We adopt the financial sentiment lexicon constructed by Hsieh et al. (2016). It was revised and expanded from Loughran and McDonald (2011) and its Chinese version (Lin 2013), which contains 1166 sentiment words, including 573 positive words and 593 negative words. Unfortunately, the dictionary contains quite a few words that are inappropriate for our work. In Hsieh et al. (2016), a revised the financial sentiment lexicon contains totally 891 words, including 369 positive words and 522 negative words, which are adopted by our work. For details on how to construct such a sentiment lexicon, please refer to (Hsieh et al. 2016).

**Sentiment Analysis**

We first retrieve sentences that include some terms in the entity lexicon. Hu and Liu (2004) observe that aspect term and sentiment term often appear close to the entity term. In our work, we follow their approach and set the distance threshold as 5. For each sentence, if the previous 5 words and the next 5 words of the entity do not involve any aspect or sentiment terms, this sentence will be discarded.

Then, we identify all sentiment words by consulting the modified financial sentiment lexicon. If a sentiment word matches some positive word in the lexicon (after considering reverse words such as "不"(No) and "無法"(cannot)), we give +1. On the other hand, if the sentiment word matches some negative word in the dictionary, we give -1. Finally, the sentiment of a news message is the average sentiment of all extracted sentiment words.
Prediction Model Construction

Before constructing a prediction model, we need to determine those features. In our work, we regard each combination of entity category and aspect category as a feature. Recall that we in our entity lexicon, there are three categories, namely industry, company, and related companies, whereas in our aspect lexicon, there are four categories, namely order, shipment, earning, and reputation. As a result, there totally 12 features that can be derived from financial news messages: industry order \( f_{IO} \), industry shipment \( f_{IS} \), industry earning \( f_{IE} \), industry evaluation \( f_{IEE} \), order of company \( f_{CO} \), shipment of company \( f_{CS} \), earning of company \( f_{CE} \), company evaluation \( f_{CEE} \), related companies’ order \( f_{CO} \), related companies’ shipment \( f_{CS} \), related companies’ earnings \( f_{CE} \), related companies’ evaluation \( f_{CEE} \). The value of each feature is a function of sentiment and count of the relevant events. Specifically, for each feature \( f \), let \( E_f \) be the set of events that describes \( f \). The feature value of \( f \) is the following:

\[
V(f) = \sum_{e \in E_f} e.\text{sentiment} \times \log(e.\text{count})
\]

Note that we take log function to reduce the effect of number of news articles about a given event. We further normalize the feature value by following the normalized function:

\[
V_{\text{normalized}} = (V_{\text{original}} - V_{\min})/(V_{\text{max}} - V_{\min}),
\]

where \( V_{\text{normalized}} \) and \( V_{\text{original}} \) are the values after and before normalization respectively, \( V_{\min} \) and \( V_{\max} \) are the minimum and maximum feature values respectively. The final normalized feature value ranges between 0 and 1.

In addition, we consider the previous revenues and apply time series prediction method for current revenue estimation. The ARIMA model (Ripley 2002) has been one of the most popular forecasting approaches for time-series data. In an ARIMA model \( (M_{ARIMA}) \), the future value of a variable is supposed to be a linear combination of past values and past errors, expressed as follows:

\[
y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}
\]

Where \( y_t \) is the actual value and \( \epsilon_t \) is the random error at time \( t \), \( \phi_i, 1 \leq i \leq p \), and \( \theta_j, 1 \leq j \leq q \), are the coefficients, \( p \) and \( q \) are integers that are often referred to as autoregressive and moving average polynomials, respectively. Basically, this method has three phases: model identification, parameter estimation and diagnostic checking. In our work, we adopt ARIMA\((1,0,1)\) model in R language, which can be represented as follows:

\[
y_t = \theta_0 + \phi_1 y_{t-1} + \epsilon_t - \theta_1 \epsilon_{t-1}
\]

Thus, in our prediction model, there are 12 features derived from financial news and one derived from previous revenue data, denoted \( f_{IS} \). Note that \( f_{IS} \) is a binary feature whose value is 1 if the revenue predicted by ARIMA is increased when compared to the last quarter’s revenue, and 0 otherwise. We can subsequently use some classification algorithms to construct the prediction model.

Evaluation

Dataset Construction

In our experiments, we select the seven Taiwan-based, major PC manufacturing companies, namely Compal, Inventec, Pegatron, Wistron, Foxconn, Gigabyte, and Quanta as our target companies. We retrieve their quarterly financial reports from Taiwan Market Observation Post System (MOPS), the system hosted by The Taiwan Stock Exchange Inc. (TWSE) & Taipei Exchange. The revenues reported in the quarterly financial reports are subsequently extracted. Our news dataset contains financial news articles crawled from online News websites about the seven companies, their industries, and related companies. We first remove stop words and the names of other similar companies, which are often mentioned together but irrelevant to our study in order to shorten the distance between relevant words. The period of these news articles is from January 2012 to December 2015, resulting in totally 40486 news articles. After processing these articles using our proposed approach, we group the news messages by the four attributes of the news message table, namely Source, Date, Entity, and Aspect,
resulting in the event table. Then the average sentiment values and the Count values for each company in each quarter are calculated and recorded. Finally, we obtained 65458 news messages and 30981 events. We have totally 112 records, i.e. 4 years * 4 quarters * 7 companies.

**Experiment Design**

We design some experiments to evaluate our prediction revenue model that utilizes both news data and previous revenue data. This model is compared to the prediction model that uses the 12 features derived from only news data and the prediction model that is based purely on ARIMA model. That is to say, we consider three prediction models: $M_{ARIMA}$, $M_{NLP}$, and $M_{ALL}$. Here $M_{ARIMA}$ refers to the pure ARIMA model that estimate the revenue trend using only the previous revenue values. $M_{NLP}$ is the model that are based on 12 textual features derived from financial news. Finally, $M_{ALL}$ is our proposed method, which involves the complete 13 features.

We observed that some companies may not have any news articles that describe their related companies for some certain seasons. In this case, the correspond records will have null values for the four sentiment features. In this case, we only retain the remaining 8 sentiment features, i.e. $f_{IO}$, $f_{IS}$, $f_{IE}$, $f_{IEE}$, $f_{CO}$, $f_{CS}$, $f_{CE}$, and $f_{CEE}$, for prediction.

**Experiment Results**

We exercised several classification methods, and here we only report the 10-fold cross validation results based on the 112 records using Sequential Minimal Optimization (SMO) in Weka, which results in the best performance. As shown in Table 5, $M_{ALL}$ achieves the best performance (in terms of accuracy, precision, recall, and F-measure), which demonstrates that the financial news data and the previous revenues both play important roles when it comes to predicting revenue trend.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{ARIMA}$</td>
<td>77.75%</td>
<td>75.40%</td>
<td>77.80%</td>
<td>76.58%</td>
</tr>
<tr>
<td>$M_{NLP}$</td>
<td>78.75%</td>
<td>76.80%</td>
<td>78.40%</td>
<td>77.59%</td>
</tr>
<tr>
<td>$M_{ALL}$</td>
<td>82.40%</td>
<td>83.60%</td>
<td>81.40%</td>
<td>82.48%</td>
</tr>
</tbody>
</table>

Our next experiment examines the news period impact. We construct two news data sets when it comes to predict the revenue trend of a company in a given season. They are 3-month data set and 6-month data set, which includes news articles about the company in the last 3 months and 6 months respectively. We intend to see if older news articles still play some roles in predicting revenue trend. As shown in Table 6, the result demonstrates that more recent news is a better predictor of revenue.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-month data set</td>
<td>78.75%</td>
<td>78.40%</td>
<td>78.40%</td>
<td>78.40%</td>
</tr>
<tr>
<td>3-month data set</td>
<td>82.40%</td>
<td>83.60%</td>
<td>81.40%</td>
<td>82.48%</td>
</tr>
</tbody>
</table>

Furthermore, we observe that in some cases the revenue difference (from the previous season) is very small and it may be inappropriate to declare the revenue rise or fall. To deal with this problem, we add one more label in addition to the original Rise and Fall: Insignificant. A revenue trend of a season is labeled Insignificant if the revenue difference of this season from the previous season is less than 0.5%. Table7 shows the confusion matrix with three labels, where $N_{R, I}$ is the number of records whose label are actually Rise but predicted as Insignificant.

| Table 6. Performance of $M_{ALL}$ Using 6-month and 3-month Data Sets
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set</td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
<td>F-Measure</td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>6-month data set</td>
<td>78.75%</td>
<td>78.40%</td>
<td>78.40%</td>
<td>78.40%</td>
</tr>
<tr>
<td>3-month data set</td>
<td>82.40%</td>
<td>83.60%</td>
<td>81.40%</td>
<td>82.48%</td>
</tr>
</tbody>
</table>

| Table 7. Confusion Matrix with Three Labels

Twenty First Pacific Asia Conference on Information Systems, Langkawi 2017
The precision and recall of label Rise are calculated by the following two equations and serve as the label performance measures. The calculation of the other labels can be similarly conducted. We adopted the macro-averaging method to present the evaluation result of this three-label classification which averages the precision and recall of every label.

\[
\text{Precision of label \textit{Rise}} = \frac{N_{R,R}}{N_{R,R} + N_{I,R} + N_{F,R}}
\]

\[
\text{Recall of label \textit{Rise}} = \frac{N_{R,R}}{N_{R,R} + N_{I,R} + N_{F,R}}
\]

### Table 8. The \textit{MALL} Method Performance of Different Classification Labels

<table>
<thead>
<tr>
<th>Measure</th>
<th>Two Labels</th>
<th>Three Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>81.40%</td>
<td>82.84%</td>
</tr>
<tr>
<td>Precision</td>
<td>83.60%</td>
<td>84.40%</td>
</tr>
</tbody>
</table>

As shown in Table 8, the prediction model based on three labels result in slightly higher precise and recalls. This experimental results suggest that marginal difference of the revenue trend should be taken into consideration. The proposed ARIMA prediction model using three labels can reach close to 85% precision values.

### Conclusion

We have proposed a revenue trend prediction approach based on financial news articles and previous revenue data. Three types of entities, namely the target company, the industry, and the related companies, and four categories of aspects, namely order, shipment, earning, and evaluation are identified as important factors. Together 12 sentiment features are used for the proposed prediction model. A lexicon-based method for determining the sentiment of each feature was proposed. In addition, one more feature computed using ARIMA based on previous revenue data is incorporated. Our experimental results using news articles of the seven Taiwan-based, major PC manufacturing companies demonstrate that both financial news articles and previous revenue data are important for accurately predicting revenue trend. The prediction model constructed using the proposed approach is able to predict revenue trend up to 84% precision. In the future work, some automatic lexicon construction methods will be considered. In addition, we plan to incorporate more sophisticated text mining techniques and other time series model in the hope to achieve even higher accuracies.

### References


