Summer 6-27-2016

PREDICTING COMPANY REVENUE TREND USING FINANCIAL NEWS

Wei-Lin Hsieh
National Sun Yat-sen University, wewaynehsieh@gmail.com

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

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

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

PREDICTING COMPANY REVENUE TREND USING FINANCIAL NEWS

Wei-Lin Hsieh, Department of Information Management, National Sun Yat-sen University, Kaohsiung, Taiwan, wewaynehsieh@gmail.com

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

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

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

Abstract

Text data analysis has found its way in many applications, and our study focuses on the financial fields. Previous studies in financial indicator prediction are mostly based on econometric models. In recent years, with the advance of text mining techniques, more and more studies employ financial news as the data source for analysis. Most studies, however, aim to predict stock prices, identify the trend of stock market, and detect company bankruptcy or company fraud. We observe that company' revenue, which can imply the company's cash flow and market share, is indeed an important financial indicator. In our study, we identify a few features that potentially impact company's revenue and further propose an approach to deriving feature values from financial news data. Specifically, we develop a lexicon-based method that involves the automatic expansion of existing financial sentiment dictionary and the aggregation of sentiment values. Preliminary experimental results show that we are able to predict the revenue trend through the news articles in the last quarter with the accuracy up to 80%.

Keywords: Text mining, Sentiment analysis, Revenue prediction, Financial news analysis, Classification.

1 INTRODUCTION

Financial news has long been regarded as an important source for understanding company's recent performance, yet investors often find it difficult to collect and digest these news articles due to their diversity, large volume, and the lack of expertise required for analysis. Recently, text mining has found its way into financial fields. However, most studies focus on predicting stock market or detecting corporate fraud. We observe that company' revenue, which imply the company's cash flow and market share, is an important financial indicator but seldom addressed by current financial texting mining researches. In this paper, we intend to fetch up this gap by proposing a method to predict company revenue from relevant financial news.

Using financial news to predict stock market is extensively researched in the last decade (e.g., see Huang et al., 2005; Zhai et al., 2007; Schumaker et al., 2012). There are even commercial services in the market, e.g., Stock Sonar, that reveal recent sentiment toward a target stock from financial news (Feldman et al., 2011). Kim (2001) suggested that profitability and revenue are the most common indicators of organization goals in the context of economic value. Rust et al. (2002) found that firms adopt primarily a revenue expansion emphasis perform better than those who emphasize on cost reduction. The relationship between revenue and news has been confirmed in Ma et al. (2009).

We identify four factors: order, shipment, earning and evaluation, and consider three scopes: the target company, the industry and the related companies, resulting in 12 features that may impact revenue. We adopt a lexicon-based approach for determining feature values from financial news and develop several methods for populating various lexicons. We retrieve 18159 news articles using four Taiwan's computer assembly companies as the target companies, namely Compal, Inventec, Pegatron, and Wistron, in 2012, 2013, and 2014. Our experiments using news data show that our approach is able to predict revenue trend up to 80% accuracy.

This paper is organized as follows. In Section 2, we present our proposed approach. Our methods for sentiment lexicon construction and sentiment determination, the major steps in our research process, are reported in Section 3. Then Section 4 shows the preliminary result of our experiments. Finally, Section 5 summarizes the paper and describes our ongoing research works.

2 THE PROCESS OF THE RESEARCH

The procedure of our approach consists of five steps. First, we determine the sources of our analysis, from which we crawl the data. Next, we define the schema of news message and event that defines how the information derived from financial news will be organized. We adopt lexicon-based approach for mining financial news. In the third step, we specify the various lexicons used in our research. The fourth step is the core step in our research, by which issues discussed in news and their sentiment are identified. Finally, we construct prediction model based on the output of the previous step, which can be used to predict revenue increase/decrease of a target company.

2.1 Data Crawling

Kinney (1971) showed that industry revenue is a good indicator to corporate and segment revenues. We thus consider news articles that describe the target company and the companies relevant to the target company, e.g., influential subsidiaries and influential joint venture, as well as the industry to which the target company belongs. The industries and the related companies are revealed in the quarterly financial reports of the target company. In addition, revenues of the target company as recorded in the reports are used for evaluating the prediction results. In our experiments, we focus on Taiwan's companies, and obtain their quarterly financial reports from Taiwan Market Observation Post System (MOPS), the system of The Taiwan Stock Exchange Inc. (TWSE) and Gre Tai Securities

Market (GTSM). The sources of financial news are crawled from several popular news websites, including China Times (CTnews), United Daily News (udn), Liberty Times (ltn), Apple Daily, and Central News Agency (CNA News). We fetch the news up to six month before the release date of a quarterly financial report for our analysis purpose. For example, suppose that the settlement date of a company's quarterly financial report is 2014/06/30, and we will take all relevant news about this company from 2014/01/01 to 2014/06/30 into account for revenue prediction.

2.2 News/Event Schema Design

We noticed that several financial news articles could be in fact about the same financial observation, usually released by the same financial agency on a particular date about some companies, which we call *events*. For example, consider the following news article we crawled from United Daily News:

*摩根士丹利科技產業分析師陳星嘉指出，筆記型電腦（NB）產業5月出貨量月增14%，優於預期，因而上調第2季NB出貨量至3,375萬台，且第3季在旺季加持下，出貨量還會再季增3%。*聯合新聞網，2014-06-13

It describes a financial observation released by Morgan Stanley (摩根士丹利), and similar reports appear in several news articles. We aim to extract messages from news articles and identify events from these messages. The schema of news messages and events are described below

2.2.1 News messages

A news message is regarded as a sentiment expression on a certain aspect of an entity, which is usually described in one or more consecutive sentences. In addition, each message is associated with some contextual information, including the source, the news agency, source announcement date, and news date. In our previous example news article, a news message will be extracted as shown in Table 1:

Table1: A news message derived from a sample news shown above

Source	Date	Entity	Aspect	Sentiment	Agency	Newsdate
摩根士丹利 (Morgan Stanley)	2014-06-13	筆記型電腦產業 (Laptop Industry)	出貨 (Shipment)	+1	聯合新聞網 (United News)	2014-06-13

2.2.2 Events

An event is defined as an announcement made by a particular source on a particular date about some aspect of an entity. We thus aggregate the information about events by the attribute values of Source, Date, Entity, and Aspect from news messages. In case a news message contains no source information, it is regarded as a distinct event with count being 1. Table 2 shows the event obtained from the previous example news, assuming that there are five news articles that report the same observation:

Table 2: An event obtained by aggregating several news messsage

Source	Date	Entity	Aspect	Sentiment	Count
摩根士丹利 (Morgan Stanley)	2014-06-13	筆記型電腦產業 (Laptop industry)	出貨 (Delivery)	+1	5

2.3 Lexicon Construction

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

2.3.1 Entity

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

Table 3: Partial entity lexicon using Compal (仁寶) as the target company

Indutry	Company	Related Companies
● 電腦及周邊設備產業	● 仁寶	● 華寶
● 筆電產業	● 2324	● 博智
● 下游硬體製造產業	● 仁寶電腦	● 智易
● PC 產業	● Compal	● 康舒
● 科技產業	● Compal Electronics	● 聯寶
● 手機代工產業	● 仁寶科技	● 樂寶

2.3.2 Aspect

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 (Kim, 1997; Macıas 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 4: Partial aspect lexicon

Order	Shipment	Earning	Reputation
● 訂單	● 出貨	● 收益	● 評價
● 接單	● 交貨	● 營收	● 大盤
● 下單	● 銷量	● 營業收入	● 買進
● 大單		● 獲利	● 賣出
● 接獲		● 營業利益	● 表現
● 蘋果單		● 利潤	● 市場

2.3.3 Source

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, 表示, and 看好 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: 里昂, 瑞信, 法說會, 花旗, 法人, 顧能, 大摩, 小摩,

DIGITIMESReaserch, 美林, 群益, 巴克萊, 瑞銀, 匯豐, 皇家, 野村, 高盛, 凱基, 麥格理, 德意志, 統一, 亨達, 英特爾, and 外資.

2.3.4 Sentiment

We adopt the financial sentiment dictionary generated by Loughran and McDonald (Loughran & McDonald, 2011) and its Chinese version (Lin, 2013). It has totally 891 words, including 369 positive words and 522 negative words.

Unfortunately, the dictionary contains quite a few words that are inappropriate to our work. For example, "銷量一夕暴增" (Sales dramatically increases overnight) is considered a sentiment word in Lin's sentiment dictionary. In our opinion, the sentiment word should be just "暴增", and "銷量" is an aspect word. Hence, we propose an automatic method to revise and expand the initial sentiment dictionary, and the detail will be described in Section 3.

2.4 Sentiment Analysis Based on Lexicons

We first retrieve sentences that include some entity term. We then follow the work of Hu and Liu (2004), which observes that aspect term and sentiment term often appear close to the entity term. In our experiment, we 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 dictionary. If a sentiment word matches some positive word in the dictionary (after considering reverse words such as "不" and "無法" as will be described in Section 3.1), 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.

2.5 Prediction Model Construction

Before constructing a prediction model, we need to determine its features (or called dimensions). In our work, we regard each entity and aspect combination as a feature, resulting in 12 features: industry order, industry shipment, industry earning, industry evaluation, order of company, shipment of company, earning of company, company evaluation, related companies' order, related companies' shipment, related companies' earnings, and related companies' evaluation. All the news related to a company's performance in a quarter is summarized as a record with 12 feature values. Referring to the news event schema, we obtain the value of each feature f by using the following equation, where E_f is the set of events that describes f, and e.sentiment and e.count represent the sentiment and message count of an event e.

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

Note that we take log function to reduce the effect of number of news messages on a given event. Our goal is to predict the revenue change, and thus the class label is 1 for revenue increase and 0 for revenue decrease. We subsequently use some classification algorithm, e.g., SVM or logistic regression to build the prediction model.

3 SENTIMENT LEXICON CONSTRUCTION

After constructing the initial lexicon, we revise sentiment words contained in the lexicon. Finally, we propose numerous rules to expand sentiment lexicon.

3.1 Initial Lexicon Construction

We have described how to construct the initial sentiment dictionary in Section 2.3.4. Negation and adverbial words play an important role in analyzing sentiment of a sentence. Negation words can be used to reverse the sentiment, and adverbial words further enhance the sentiment of the word. We build negation and adverbial words lexicon by following the work of (Zagibalov & Carroll, 2008), which uses the most common negation words such as 不(bu), 不會(buhui), 沒有(meiyou), 擺脫(baituo), and 避免(bimian), and the most common adverbials such as 很(hen), 非常(feichang), 最(zui), and 比較(bijiao). Furthermore, we also found that 走(zou),and 看(kan) are frequently modified sentiment words in the financial field (e.g. 走高, 看好), so we add them to our adverbial lexicon.

We also prepare a stop word list which is obtained from Word List with Accumulated Word Frequency in Sinica Corpus 3.0 (Sinica Corpus, 2016) with some manual adjustment. For example, we remove one of our negation word 沒有(meiyou) from the stop word list.

3.2 Revising Sentiment Dictionary

Based on the initial sentiment dictionary, we formulate some rules to revise it. In the beginning, we found some unreasonable words which appear in both positive and negative lexicons, so we manually determine their appropriate polarity or remove them if their polarities are not clear in the financial domain.

```
1   Sentiment_Revision(S: a set of <word, value> pairs): a revised sentiment dictionary
2   {
3     S' = ∅;
4     For each <word, value> in S do
5       Remove Punctuation(word);
6       Remove Non-Chinese(word);
7       Remove De(word);
8       switch (Length(word))
9         case '1' :
10          Uni_Word(word, value);
11        case '2' :
12          Bi_Word(word, value);
13        default :
14          More_Word(word, value);
15      if value != 0
16        Remove Duplicate(word);
17    return S';
18  }
```

Figure 1. The Main Pseudocode of Modifying Sentiment Words

After manually adjusting unreasonable words, we draw up an algorithm to revise the dictionary by the following steps. The main pseudocode is shown in Figure 1. Firstly, punctuations are removed. For

example 漲幅「第一名」 will be modified into 漲幅第一名. Then, non-Chinese words (e.g., No.1) are removed. Third, 的 (de) is removed from the terms ending with 的 (de) because they carry no special meanings other than adjectives. For example, 弱的 will be modified to 弱, and 下跌的 will be modified to 下跌. Finally, we extract core sentiment words from terms in the sentiment dictionary, which is the core function of revising sentiment dictionary and will be described in more detail in the following.

For each sentiment word (after the first three steps), if it never occurs in our news corpus, it will be deleted, otherwise the following procedure is performed. If the word length is one character, the function Uni_Word(*word*, *value*) is performed to check if *word* exists in entity lexicon, aspect lexicon, source lexicon, source lexicon, negation lexicon, adverbial lexicon, or stop word list as described previously. If so, (*word*, *value*) will be deleted from the sentiment lexicon; otherwise, it will be kept.

If the word length is two characters, The function Bi_Word(*word*, *value*) is performed. We identify a sentence in our corpus that contain *word* and use CKIP, a popular Chinese word segmentation tool (CKIP, 2016), to segment the sentence. If the entire sentiment word is treated as a component word by CKIP, we will keep the sentiment word. Otherwise, if *word* is separated into two word, w_1 and w_2, and w_1 appears in the negation lexicon, we will keep w_2 and reverse its sentiment value. For example, ‘不良’ (buliang) is shown in the initial sentiment lexicon as a negative word. CKIP splits it into ‘不’ and ‘良’, and ‘良’ will be added to the sentiment lexicon as a positive word. If w_1 appears in the adverbial lexicon, w_2 is regarded as a sentiment word with the same sentiment value. For example, ‘最優’ (zuiyou) is segmented into ‘最’ and ‘優’, and ‘優’ will be insert into the positive sentiment lexicon. Similar procedure is used to handle the case where there are no less than three characters, i.e., More_Word(). For space limitation, we omit it from the paper.

3.3 Expanding Sentiment Dictionary

We observe that some popular sentiment word such as ‘看好’ does not appear in our revised sentiment lexicon. However, often times, it includes some character appears in our sentiment lexicon, e.g., ‘好’. Firstly, we identify all sentences in our corpus that contain some single-character sentiment words in our corpus. Then we segment each sentence by CKIP and retrieve the component words that contain some single-character sentiment word. Only expanded words with POS tags intransitive verb (Vi) or adjective (A) are preserved, and their polarity will be the same as the containing single-character sentiment word. For example, an expanded word ‘尚佳’ will be added to the sentiment lexicon with polarity being positive, the same as ‘佳’. An exception is ‘開高走低’, which contains both positive sentiment word ‘高’ and negative sentiment word ‘低’. In this case, by consulting the original sentiment dictionary, we find ‘開高走低’ is actually a negative term.

After revising original sentiment dictionary and expanding modified sentiment dictionary, our final sentiment dictionary contains 1166 sentiment words, including 573 positive words and 593 negative words.

4 PRELIMINARY EVALUATION

Our news dataset contains financial news articles crawled from online News websites for the four companies (Compal, Inventec, Pegatron, and Wistron), 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. After processing using the approach described in Section 3, we obtain 32379 messages 13309 events in our database.

We used the method of Sequential Minimal Optimization (SMO) which is based on Support Vector Machine (SVM) in Weka (Waikato Environment for Knowledge Analysis) to construct the prediction

model. We adopt 10-fold cross validation to analyse the three methods. The result shows that the proposed approach may achieve accuracy up to 80.65%.

5 CONCLUSION

Our classification framework identifies four factors, namely order, shipment, earning, and evaluation, and considers the target company, the industry, and the related companies. We adopted a lexicon-based approach for identifying feature values from financial news and developed several methods for populating various lexicons. Our experiments using news data in the last quarter of four computer assembly companies show that our model is able to predict revenue up to 80% accuracy. This framework can be generally applied to various industries to predict their revenues or other important financial indicators through financial news.

Because the principle of organizing financial statements in Taiwan has been significantly changed since 2012, we can only collect three years of financial reports, resulting in only 12 records for each company. Our ongoing work includes involving more companies of the same industry for our analysis and conducting more experiments. In addition, the accuracy of our lexicon construction is heavily affected by the NLP tool. In the future, the development of better NLP tools for Chinese could further increase the accuracy of our proposed approach.

References

CKIP (2016), Chinese Knowledge and Information Processing, available at http://rocling.iis.sinica.edu.tw/CKIP/engversion/index.htm.

Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011). The stock sonar—sentiment analysis of stocks based on a hybrid approach. Paper presented at the Twenty-Third IAAI Conference.

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513-2522.

Kim, Y. (1997). Measuring efficiency: The economic impact model of reputation. Paper presented at the annual conference of the Public Relations Society of America, Nashville, TN.

Kim, Y. (2001). Measuring the economic value of public relations. Journal of Public Relations Research, 13(1), 3-26.

Kinney, W. R. (1971). Predicting earnings: entity versus subentity data. Journal of Accounting Research, 127-136.

Lin, I.-H. (2013). Creating and Verifying Sentiment Dictionary of Finance and Economics via Financial News, master thesis, National Taiwan University.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10- Ks. The Journal of Finance, 66(1), 35-65.

Macıas, M., Guitart, J., Center, B. S., & Girona, J. (2008). Influence of reputation in revenue of grid service providers. Paper presented at the 2nd International Workshop on High Performance Grid Middleware (HiPerGRID 2008).

Rust, R. T., Moorman, C., & Dickson, P. R. (2002). Getting return on quality: revenue expansion, cost reduction, or both? Journal of marketing, 66(4), 7-24.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems, 53(3), 458-464.

Sinica Corpus (2016), Sinica Corpus 3.0, available at http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm.

Zhai, Y., Hsu, A., & Halgamuge, S. K. (2007). Combining news and technical indicators in daily stock price trends prediction Advances in Neural Networks, Fourth International Symposium on Neural Networks (ISNN 2007).

Ma, Z., Shen, O. R. L., Pant, G. (2009). Discovering company revenue relations from news: A network approach. Decision Support Systems, 47(4), 408-414.