Analysis of Stock Market using Text Mining and Natural Language Processing

Sheikh Shaugat Abdullah, Mohammad Saiedur Rahaman and Mohammad Saidur Rahman
Department of Computer Science
American International University-Bangladesh, Dhaka, Bangladesh
{shaugat, saied, saidur}@aiub.edu

Abstract— **Stock market has become one of the major components of economy not only in developed countries but also in third world developing countries. Making decision in stock market is not really easy because a lot of factors are involved with every choice we make. Therefore, a lot of analysis is required to make an optimal move on stock market which may involve price trend, market's nature, company's stability, different news and rumors about stocks etc. The objective of this study is to extract fundamental information from relevant news sources and use them to analyze or sometimes forecast the stock market from the common investor's viewpoint. We surveyed the existing business text mining researches and proposed a framework that uses our text parser and analyzer algorithm with an open source natural language processing tool to analyze (machine learning and text mining), retrieve (natural language processing), forecast (compare with historic data) investment decisions from any text data source on stock market. For our research we used the data of Dhaka Stock Exchange (DSE), capital market of Bangladesh.**

Index Terms— Business Text Mining, Stock market forecasting, Rule extraction, Information Gathering

I. INTRODUCTION

Making investment decisions in stock market is risky sometimes because it is the fastest and also easiest way of making money as well as losing money. Therefore, investing on stock market needs careful planning with deep analysis which now a days is possible using advanced technologies with large computational power, neural network, relational database etc.

Stock market analysis can be separated into two categories. One is Technical and another one is fundamental. Technical analysis looks at the price trend of a security and uses this data to forecast its future price movements where fundamental analysis, on the other hand, looks at economic factors, known as fundamentals. Though both type of analysis are important but technicians believe that all the information they need about a stock can be found in its charts and therefore technical traders, believe there is no reason to analyze a company's fundamentals because these are all accounted for in the stock's price [7]. However to understand the long term performance of a stock, to estimate the risk factors involve d with an investment, to understand the entry and exit point in a volatile stock market, fundamental analysis is mandatory and news, records, rumors about stocks can provide necessary input for fundamental analysis.

In this research, we applied text mining techniques on different news and articles published by legitimate companies on internet to extract high quality information which ultimately made it possible to analyze, decide and also update our database for other type of analysis. We proposed a framework for information gathering and processing where we used a natural language processing tool with our suggested algorithms.

This research is divided into four sections where the first section provides a basic summery of works that are related to our research. In the second section, the research methodology is described and in the third section we illustrated our proposed framework with a few examples and analysis result in the following section. Finally we have shown how news can lead to a profitable decision in the last section.

II. RELATED WORKS

In 2002, Antonina Kloptchenko, Tomas Eklund, Barbro Back, Jonas Karlsson, Hannu Vanharanta and Ari Visa used data and text mining techniques to study hidden patterns about the future financial performance of companies from the quantitative and qualitative parts of their financial reports. Because from their point of view, annual reports are an important medium for the company's communication with its stakeholders [1]. However, their text mining approach can only be applied on annual reports of a company, not web sites or news achieve.

Later on in 2006, Marc-André Mittermayer and Gerhard F. Knolmayer performed a survey on text mining systems for predicting market response to news where they briefly described and compared eight existing prototypes in chronological order [2]. But none of the techniques showed mechanism of retrieving decisions from reumors using natural language processing. On the same year Robert P. Schumaker and Hsinchun Chen introduced a predictive machine learning approach AZFinText System for financial news articles analysis using several different textual representations like Bag of Words, Noun Phrases, Named Entities and investigated 9,211 financial news articles and 10,259,042 stock quotes during a five week period [3]. Their technique was very

978-1-4799-0400-6/13/$31.00 ©2013 IEEE

efficient and rapid but they did not show any decision making process (based on historic data) from the retrieved information.

In 2007, Vatsal H. Shah and Dr. Mehryar Mohri presented a machine learning techniques for stock prediction where they particularly discussed the application of Support Vector Machines, Linear Regression, prediction using decision stumps, expert weighting and online learning in detail along with the benefits and pitfalls of each method [6]. The machine learning techniques they discussed were slightly focused on text and language processing.

Nan Li and Desheng Dash Wu presented text sentiment analysis, also called emotional polarity computation where they studied online forums hotspot detection and forecast using emotion analysis and text mining approaches in early 2009 [4]. Their technique showed a totally a different idea of forecasting compare to other techniques.

Afterward in 2009, Xiangyu Tang and Chunyu Yang, Jie Zhou proposed an algorithm for stock price forecasting by combining news mining and time series analysis [5]. The idea of using time series analysis with text mining was discussed in their work.

III. Methodology

In the related work section we discussed some works from where we started our research. In this paper we showed how information can be retrieved from textual data, how the retrieved information can be directed to a forecasting decision with the help of historic data, how our algorithm can train itself as it process different types of news, how textual data with no particular pattern can be processed using natural language process tool. The idea we discussed in this research is different than others because we considered and combined information parsing technique from two types of text data (patterned or scattered), we considered historic data before making decision and we showed rumors are a valuable data source.

Analyzing the stock market requires analysis to textual data related to the market as well as analysis with numerical data. The textual data may be news, articles, notice, announcement or sometimes rumors about some stocks or the whole stock market. On the other hand, numerical data actually represents price, volume, turnover, number of outstanding shares etc. of the enlisted companies. However, both type of information is equally important for making any decision on stock market. The challenge is to deal with textual data as they can be written in different format and these data can be collected from different sources. To overcome this problem we proposed a framework to capture textual data related to stock market from diverse sources, categorize them and extract related information which helps us to make investment decision. Moreover, we are storing that information at the same time to enrich our database for other type of analysis. The figure 1 shows an overview of our methodology.

In our proposed system, based on the input of the user we proposed him/her the set of stocks that is suitable for their choice. To prepare this set, we considered different fundamental factors like EPS, P/E, beta, and co-relation, standard deviation etc that we have captured, extracted or derived directly or indirectly during information parsing process. Therefore, the selection procedure of stock actually depends how much risk an investor is willing to take in the market.

Though, these fundamental factors only help us to screen out and prepare a safe list of stocks depending on risk factor setting but to get the appropriate time of investment on these stocks, we need to consider price trend related factors which refers to technical analysis.

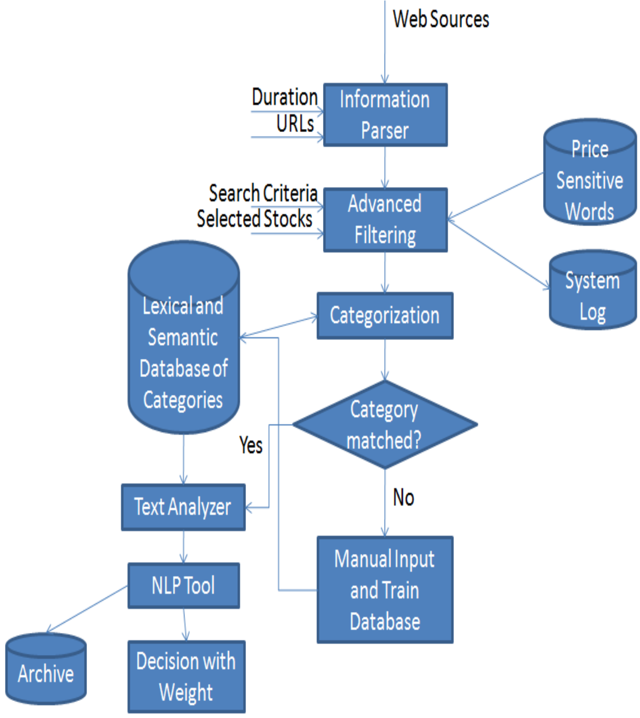


Fig. 1. Overview of textual data capturing, analyzing, categorizing and information extracting process.

IV. Proposed Textual Analyzer Framework

After In the previous section we saw technical indicators only work on the price of a security, price trends considering a very short period of time. But whatever the price trend revels, before investing on any stock one must consider the fundamental position of a security like:

How risky (weighted risk) the stock is?

How it is co-related with the market?

How is the financial report of that company?

The stock belongs to which category of the market?

What is the weight (outstanding share) of that stock compare to whole market? etc.

There are a lot of questions and more are arising with time, in our mind whenever we try to take any investment decision. Nevertheless, it is very difficult for the investors to keep track of all these factors, for all the stocks, for all the time .Because there may be thousands of companies enlisted in a stock market, millions of news sources with billions of news around. Therefore, it is impractical for an individual, manageable for a research farm but very much possible for a machine.

In our proposed framework information can be collected from different sources where information parsing is allowed. For official information like Bonus or Dividend or changes on outstanding shares or merge of companies or split of a company or Director Buy/Sell etc. should only be collected from authenticated sources and therefore collected from that specific company's website or from stock exchange server. On the other hand, chitchat or rumor which may or may not be important is collected from blogs, facebook, twitter and whatnot though the user of the system must input the ranges of URL where to surf. Moreover user must also provide duration input to make sure how frequently the sites will be checked.

In subsequent layer, we applied advance filtering on recently parsed data to ensure whether the data is relevant or not. To do so, users must provide a list of stocks where they are interested in and keywords that implies; on which factors of those stocks are important for them. We used a database of keywords for filtering and a system log is always maintained for all the parsed data for future reference.

While the parsed data passed the filtering process as a suitable candidate for valuable information, we tried to categorize the news. Data form same category share common set of keywords with common semantics. Thus we maintained lexical and semantic resources of each category in a database.

If any of the category criteria is not matched with the parsed data, then the system will throw an exception and wait for manual input. When the required input is provided by user, the system will train itself by updating its database according to manual input and therefore in future if any news belongs to same category arrives, it can categorize without any interruption.

On the other hand, if the category is matched then the system already knows how to analyze any data from that category based on the database of lexical and semantic resource thought user can set a precision factor depends on what percentage of the parsed data should be matched with the resources of that category.

Finally, we used an open source natural language processing tool for extracting information or in some cases decisions from analyzed text and stored them in achieve for further analysis and calculations. For our analysis we used Apache OpenNLP which is a Java based machine learning toolkit for natural language processing. To extract any decision from parsed data, we compared the latest information with database to identify the impact of the news on market. In case of rumor, the system provides an output considering the fundamental factors of the concerned stock. So, if the fundamental factors of that stock support the rumor, then weight factor of that decision will be high and this may lead us to ignore technical input for some cases.

V. Analysis and Experiment Results

In this section, we will show various example of our analysis and offer some results of our experiment. The data from same category of different stocks shares some general attributes and we used that pattern to train our algorithm so that we can retrieve information from any data of that category. Moreover, if the pattern changes with time, the system will prepare itself by changing the keywords and semantics in database.

For instance the following news belongs to same category thought their origin is not same [8]:

News1:

The Board of Directors has recommended **cash dividend** @ 5% and **stock dividend** @ 20% *for the year ended on* December 31, 2011. *Date of AGM:* 06.05.2012, Time: 11:00 AM, Venue: Bangabandhu International Conference Centre, Agargaon, Sher-E-Bangla Nagar, Dhaka. **Record date:** 01.03.2012. The Company has reported Consolidated **EPS** of Tk. 3.78, Consolidated **NAV per share of** Tk. 40.74 and Consolidated **NOCFPS of** Tk. 36.10 for the year ended on December 31, 2011. The Company has also reported EPS of Tk. 3.60, NAV per share of Tk. 40.29 and NOCFPS of Tk. 34.15 for the year ended on December 31, 2011.

News2:

The Board of Directors has recommended **stock dividend** @ 21% *for the year ended on* December 31, 2011. *Date of AGM:* 30.06.2012, Time: 10:30 AM, Venue: Beximco Industrial Park, Sarabo, Kashimpur, Gazipur. **Record date:** 21.05.2012. The Company has also reported EPS of Tk. 4.76, **NAV per share of** Tk. 68.03 and **NOCFPS of** Tk. 6.26 for the year ended on December 31, 2011.

News3:

The Board of Directors has recommended **cash dividend** @ 10% and **stock dividend** @ 15% *for the year ended on* December 31, 2010. *Date of AGM:* 13.07.11, Time: 11:00 AM, Venue: Bangabandhu International Conference Centre, Agargaon, Sher-E-Bangla Nagar, Dhaka. **Record date:** 19.05.11. The Bank has reported consolidated profit after tax of Tk. 3,989.52 million, consolidated **EPS** Tk. 124.47, consolidated **NAV per share of** Tk. 441.36 and consolidated **NOCFPS of** Tk. 44.78 for the year ended on 31.12.10. The Bank has also reported profit after tax of Tk. 3,696.02 million, EPS Tk. 115.31, NAV per share of Tk. 432.61 and NOCFPS of Tk. 202.06 for the year ended on 31.12.10.

In the above example, all the news is actually conveying the same message which made them a group. Our proposed system will detect the type of the news form italic keywords. However to confirm the category of the news the bold marked text will be checked which also give us a marker where to read the required information as well. Finally the underlined text is actually the information we are interested in from this news. Therefore the summery of the information we have retrieved:

TABLE I. SUMMARY OF INFORMATION THAT WE HAVE EXTRACTED FROM THE ABOVE EXAMPLE

Company Name	Record Date	Cash Dividend	Stock Dividend	EPS
xxxx	01-03-2012	5%	20%	3.78
xxxx	21-05-2012	#N/A	21%	4.76
xxxx	19-05-2011	10%	15%	124.47

The above information will be updated on relevant database. Another example with a different category is given below:

News1:

Mr. yyyy, *one of the Directors of the Company*, *has expressed* his *intention* to **buy** 24,35,000 shares of the Company at prevailing market price through Stock Exchange **within** next 30 working days.

News2:

Mr. yyyy, *one of the Directors of the Company*, has further reported that *he has completed* his **buy** of 24,35,000 shares of the Company at prevailing market price through Stock Exchange as announced earlier.

News3:

Mr. zzzz, *one of the Directors of the Company* as per Company's report dated 09.05.2012, *has expressed* his *intention* to **buy** 4,05,000 shares of the Company at prevailing market price through Stock Exchange **within** next 30 working days.

News4:

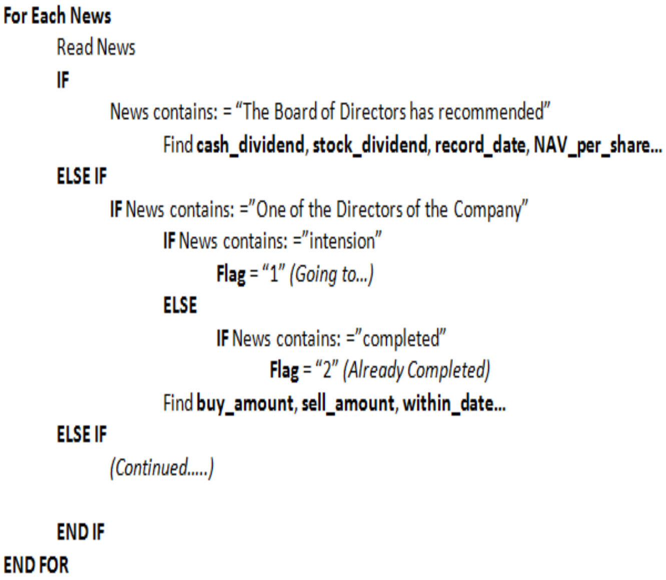
Mrs. zzzz, *one of the Directors of the Company*, *has expressed* her *intention* to **sell** 2,00,000 shares out of her total holding of 17,33,000 shares of the Company at prevailing market price through Stock Exchange **within** October 31, 2012.

For the above example, there are actually two categories of news, one is representing a stakeholder of the company has showed intention to Buy/Sell some shares and the other one actually represents the stakeholder actually completed the intention of buying/selling. Thus, News1, News3 and News4 of example two contains the common keywords and semantics which made them remain in same set. As stated before, the italic keywords are used to verify the message type and the bold words are used to cross check the previous decision and also to get a point where to look for our data (figure-2). Again the underlined sections are actually information that we have retrieved form that parsed news.

TABLE II. SUMMARY OF BOTH TYPE OF INFORMATION THAT WE HAVE EXTRACTED FROM THE EXAMPLE

Company Name	Number of Shares	Decision	Range	Status
xxxx	24,35,000	Buy	30 working days	Intention
xxxx	24,35,000	Buy	#N/A	Completed
xxxx	4,05,000	Buy	30 working days	Intention
xxxx	2,00,000	Sell	October 31, 2012	Intention

As we have used the data of Dhaka Stock Exchange for our analysis, therefore, we applied our algorithm on all the existing news to test the applicability of our algorithm. Most of the cases the algorithm successfully categorized the news data and added information on database although due to changed nature of some news, manual interruptions were required for some cases.



```
For Each News
    Read News
    IF
        News contains: = "The Board of Directors has recommended"
            Find cash_dividend, stock_dividend, record_date, NAV_per_share...
    ELSE IF
        IF News contains: ="One of the Directors of the Company"
            IF News contains: ="intension"
                Flag = "1" (Going to...)
            ELSE
                IF News contains: ="completed"
                    Flag = "2" (Already Completed)
            Find buy_amount, sell_amount, within_date...
    ELSE IF
        (Continued.....)

    END IF
END FOR
```

Fig. 2. Part of pseudo code for analyzing text for the above two examples.

Nevertheless once the algorithm understands the new pattern; it tries to apply that technique on other data. Figure 3 presents a small part of the database that we have retrieved by applying our text analyzer algorithm on news archive of Dhaka Stock Exchange. In the table, some of the fields can be directly parsed and some of them are derived.

VI. INFORMATION BASED DECISION

Sometimes information or news is more important than any other type of analysis. For example, if the Government forces the directors of a company to buy a large quantity shares within a limited timeframe then anyone can understand whatever the analysis(for instance technical) result is the price of that stock is going to increase. So, people should buy that stock at that point. Now, it may happen, technical analysis shows "sell" decision at that point which is totally not realistic because it only considers the price trend. To make the prediction result more realistic, news, information or rumor about individual stock or any sector of the market or the whole stock market need to be considered.

Therefore, some of the news can lead us to a clue that may help us to take decision. For example, for any stock if the stock dividend of current year in almost double of previous year, then it will be positive news for the shareholders and sometimes by collecting some advance news people tries to buy that share. So the price of that share increases. On the other hand, if there is no dividend whereas there was dividend provided in previous year, then the news will have negative impact on the shareholders for that stock.

Moreover, price of the stock are generally decrease after the record date as people tries to realize their gain, so the stock price decreases most of the time. The same fact is true for director buy/sell announcement. The following table shows offered stock dividend percentage for four consecutive years with their probable impact on market due to their changes compare to previous year.

TABLE III. STOCK DIVIDEND OF ABBANK FOR FOUR YEARS STARTED FROM 2006

Year	Percentage	Impact
2006	10%	#N/A
2007	30%	Positive (200% Increment)
2008	200%	Positive (566% Increment)
2009	15%	Negative (92.5% Decrement)

However, News from unauthentic sources or rumors does not follow any pattern of any particular format and therefore cannot be categorized like news form authentic sources. But sometimes rumor can provide valuable information though it is very risky and unauthentic. Since this type of news do not follow any pattern so used natural language processing tool to pull out information from them. We also tried to set a weight of the rumors that represents the significance of the decision retrieved.

The weight of the decision can be calculated from the fundamental background of the stock for which the decision was made. If the fundament facts of that company do not support the information from rumor then the algorithm will put less weight and vice versa.

VII. CONCLUSION AND FUTURE WORK

This research proposes a new data processing framework that takes text from different sources as input where the source may be authentic or unauthentic. For authentic sources like company web source or stock exchange database, the information is retrieved using a text analyzer algorithm and stored in database. However information from unauthentic sources are passed to a natural language processing tool to extract as much information as possible and therefore set a rank or weight on that information by comparing with historical data. Though we used the data of Dhaka Stock Exchange (DSE) for our analysis, this technique is applicable on any other stock exchange as well [8].

Any news can lead us to decision by comparing it with chronological data which may or may not be true depending on other factors like price trends. Therefore, in future we have a plan to incorporate the technical analysis with our proposed framework to make the investment decision more precise and pragmatic.

REFERENCES

[1] Antonina Kloptchenko, Tomas Eklund, Barbro Back, Jonas Karlsson, Hannu Vanharanta and Ari Visa, Combining Data And Text Mining Techniques For Analyzing Financial Reports: Eighth Americas Conference on Information Systems, 2002, pp. 20-28.

[2] Marc-André Mittermayer and Gerhard F. Knolmayer, Text Mining Systems for Market Response to News: A Survey: Working Paper No 184, Institute of Information Systems, University of Bern, August 2006.

[3] Robert P. Schumaker and Hsinchun Chen, Textual Analysis of Stock Market Prediction Using Financial News Articles: 12th Americas Conference on Information Systems (AMCIS), 2006.

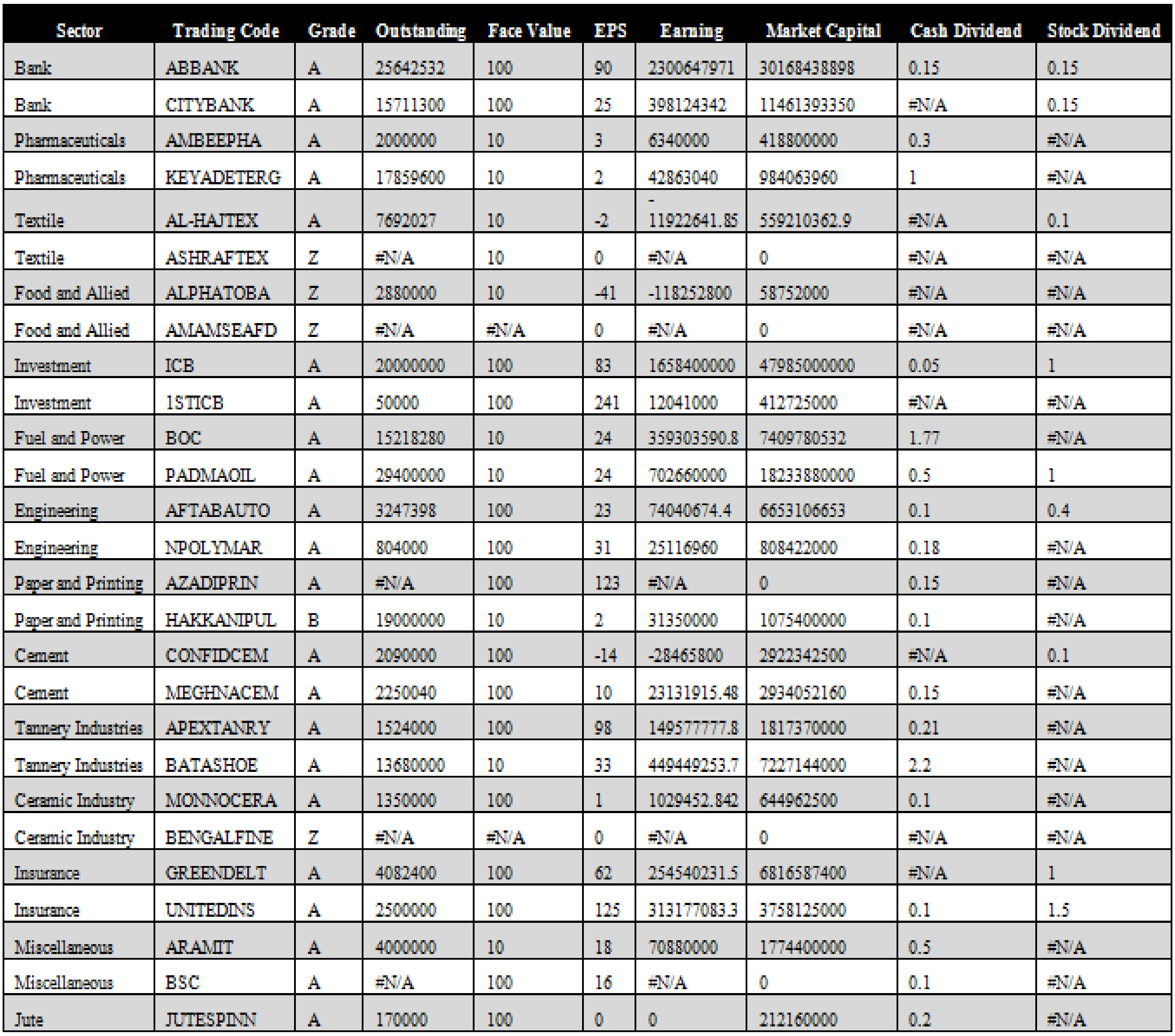
[4] Nan Li and Desheng Dash Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast: Decision Support Systems 48, 2010, pp. 354–368.

[5] Xiangyu Tang and Chunyu Yang, Jie Zhou Stock Price Forecasting by Combining News Mining and Time Series Analysis: 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, ISSN 978-0-7695-3801-3/09, pp. 279-282.

[6] Shah V. H. and Dr. Mohri M. Machine Learning Techniques for Stock Prediction: Foundations of Machine Learning: Spring 2007, Courant Institute of Mathematical Science, New York University.

[7] Technical Analysis: Fundamental Vs. Technical Analysis: www.investopedia.com Retrieved on: 14/12/2012.

[8] Dhaka Stock Exchange (DSE) News Achieve: www.dsebd.org Retrieved on: 10/11/2012.



Sector	Trading Code	Grade	Outstanding	Face Value	EPS	Earning	Market Capital	Cash Dividend	Stock Dividend
Bank	ABBANK	A	25642532	100	90	2300647971	30168438898	0.15	0.15
Bank	CITYBANK	A	15711300	100	25	398124342	11461393350	#N/A	0.15
Pharmaceuticals	AMBEEPHA	A	2000000	10	3	6340000	418800000	0.3	#N/A
Pharmaceuticals	KEYADETERG	A	17859600	10	2	42863040	984063960	1	#N/A
Textile	AL-HAJTEX	A	7692027	10	-2	-11922641.85	559210362.9	#N/A	0.1
Textile	ASHRAFTEX	Z	#N/A	10	0	#N/A	0	#N/A	#N/A
Food and Allied	ALPHATOBA	Z	2880000	10	-41	-118252800	58752000	#N/A	#N/A
Food and Allied	AMAMSEAFD	Z	#N/A	#N/A	0	#N/A	0	#N/A	#N/A
Investment	ICB	A	20000000	100	83	1658400000	47985000000	0.05	1
Investment	1STICB	A	50000	100	241	12041000	412725000	#N/A	#N/A
Fuel and Power	BOC	A	15218280	10	24	359303590.8	7409780532	1.77	#N/A
Fuel and Power	PADMAOIL	A	29400000	10	24	702660000	18233880000	0.5	1
Engineering	AFTABAUTO	A	3247398	100	23	74040674.4	6653106653	0.1	0.4
Engineering	NPOLYMAR	A	804000	100	31	25116960	808422000	0.18	#N/A
Paper and Printing	AZADIPRIN	A	#N/A	100	123	#N/A	0	0.15	#N/A
Paper and Printing	HAKKANIPUL	B	19000000	10	2	31350000	1075400000	0.1	#N/A
Cement	CONFIDCEM	A	2090000	100	-14	-28465800	2922342500	#N/A	0.1
Cement	MEGHNACEM	A	2250040	100	10	23131915.48	2934052160	0.15	#N/A
Tannery Industries	APEXTANRY	A	1524000	100	98	149577777.8	1817370000	0.21	#N/A
Tannery Industries	BATASHOE	A	13680000	10	33	449449253.7	7227144000	2.2	#N/A
Ceramic Industry	MONNOCERA	A	1350000	100	1	1029452.842	644962500	0.1	#N/A
Ceramic Industry	BENGALFINE	Z	#N/A	#N/A	0	#N/A	0	#N/A	#N/A
Insurance	GREENDELT	A	4082400	100	62	254540231.5	6816587400	#N/A	1
Insurance	UNITEDINS	A	2500000	100	125	313177083.3	3758125000	0.1	1.5
Miscellaneous	ARAMIT	A	4000000	10	18	70880000	1774400000	0.5	#N/A
Miscellaneous	BSC	A	#N/A	100	16	#N/A	0	0.1	#N/A
Jute	JUTESPINN	A	170000	100	0	0	212160000	0.2	#N/A

Fig. 3. Part of the database retrieved by applying our text analyzer algorithm on news archive of Dhaka Stock Exchange.