

CLASSIFICATION OF MARKET NEWS AND PREDICTION OF MARKET TRENDS

P. Kroha and R. Nienhold

Department of Computer Science, University of Technology, Strasse der Nationen 62, 09111 Chemnitz, Germany

Keywords: Text classification, Stock exchange news, Market trends.

Abstract: In this contribution we present our results concerning the influence of news on market trends. We processed stock news with a grammar-driven classification. We found some potentialities of market trend prediction and present promising experimental results. As a main result we present the fact that there are two points (altogether only two) representing the minimum of good news/bad news relation for two long-term trends breaking points (altogether only two) during the last 10 years and that both of these points representing news appear in both cases about six months before the long-term trend of markets changed.

1 INTRODUCTION

Text mining is a promising new field combining methods of linguistics, computer science, artificial intelligence, and mathematics. The goal is to extract, represent, and use the meaning of texts. A subset of text mining is text classification. Text classification is easier because the topic is not the meaning of a text but the similarity of textual objects. Classification methods are known from other fields of applications. The main idea is that we have to find specific features of objects to be classified that can be used to define classes and to control the process of assigning an object to a class.

One practical application of text classification is the classification of short messages, e.g. advertisements. We considered it interesting to classify news concerning markets because we can investigate the influence of the obtained classes and their cardinalities on market prediction after the classification.

There are some thousands of daily news available that can influence markets. Nobody can read them all. Usually, analysts are specialized and read only news of some branch, published by some respectable agency.

Most investors believe that markets are driven by news so that they need to read them and use them for their investment decision. Because of that, news about economic and political events that concern markets and stock exchange situation are a very well-selling article. There are profitable agencies selling news subscriptions of any possible kind.

Theoretically, this concept is controversial or at least disputed. The currently accepted theory of efficient markets (Fama, 1970) states that the information contained in any public news story is already contained in stock prices so that it cannot bring a profit anymore. As we started our investigation we stated a hypothesis based on the theory of efficient markets. Our hypothesis was that when news only describe the existing situation, i.e. when the information obtained in news is already contained in stock prices then news cannot have any prediction features.

However, there is also the theory of inefficient markets (Shleifer, 2000) that has interesting arguments, too. Practically, investors hope that markets are inefficient and believe that they can use various strategies how to exploit the news obtained. Some of them are buying on good news, some of them are selling on good news. If the majority of investors wants to buy, the up-trend will be generated. If the majority of investors wants to sell, the down-trend will be generated.

The process is complex because the market context changes very chaotically. Many investors do not have a stable behavior and they do not use fixed rules but often only an intuition. The situation which is necessary for the stock exchange to work is that at any time point there is a group of investors thinking that this is the right point to sell and at the same time point there is another group of investors thinking that this is the best time point to buy. In other cases, the stock exchange would not work.

Our motivation was to investigate the influence of

news on long-term market trends using different techniques of text classification. We continued the research presented in (Kroha and Baeza-Yates, 2005), (Kroha et al., 2006), (Kroha and Reichel, 2007), (Kroha et al., 2007) that resulted in the conclusion that the quotient between good and bad news starts to recover some months sooner before the markets start to grow up again.

In (Kroha and Reichel, 2007), we introduced our grammar-driven text classification method for market news in English and found that the curve representing the quotient between positive and negative news starts to grow about some months before the markets start to grow. Because we investigate long-term trends and because we have our data starting in 1999 we found only one such situation described above.

Ever since, the prime crisis has come and there is another down trend (2007 - 2009) and another breaking point (March 2009) in markets.

In the investigation described in this paper, we want to answer the following questions:

- How easy or difficult it is to construct a grammar for our grammar-driven classification of textual market news when it should be done for German language (complex declination, conjugation, irregular plurals)? We were trying to answer it for German language because we have about 554.000 news available.
- When we change not only the language of news but also the time interval do we get similar results compared with our previous? This means, we were curious whether the dependences found for english market news from the time interval 1999 - 2006 can be observed in german market news from the time interval 1999 - 2009, too.

We used the grammar-driven method like we did in (Kroha and Reichel, 2007) but we constructed a grammar for classification of news written in German. Finally, we found that the conclusion from analysing market news in English from 1999 - 2006 can be confirmed through experiments with market news in German from 1999 - 2009. As we will show, the news indicator (good news/bad news) processed by our grammar-driven text classification method seems to have interesting features that could be used for forecasting because the news indicator changes its trend some months before the market changes its trend.

The rest of the paper is organized as follows. In Section 2 we discuss related work. In Section 3 we briefly explain why we want to use a grammar-driven method. In Section 4 we take a look to the implemented system which proceed all messages. Section 5 describes our results in comparison to the DAX stock

index. In the last section we present some measurement results and conclude.

2 RELATED WORK

In related papers, the approach to classification of market news is similar to the approach to document relevance. Experts construct a set of keywords which they think are important for moving markets. The occurrences of such a fixed set of several hundreds of keywords will be counted in every message. The counts are then transformed into weights. Finally, the weights are the input into a prediction engine, which forecasts which class the analyzed message should be assigned to.

In (Nahm and Mooney, 2001) and (Nahm, 2002), a small number of documents was manually annotated (we can say indexed) and the obtained index, i.e. a set of keywords, will be induced to a large body of text to construct a large structured database for data mining. The authors work with documents containing job posting templates. A similar procedure can be found in (Macskassy and Provost, 2001). The key to his approach is the user's specification to label historical documents. These data then form a training corpus to which inductive algorithms will be applied to build a text classifier.

In (Lavrenko et al., 2000), a set of news is correlated with each trend. The goal is to learn a language model correlated with the trend and use it for prediction. A language model determines the statistics of word usage patterns among the news in the training set. Once a language model has been learned for every trend, a stream of incoming news can be monitored and it can be estimated which of the known trend models is most likely to generate the story. Compared to our investigation, there are two differences. One difference is that Lavrenko uses his models of trends and corresponding news only for day trading. The weak point of this approach is that it is not clear how quickly the market responds to news releases. The next difference is that our grammar-driven method respects the structure of a sentence that can have a fundamental influence on the meaning of the sentence.

In our previous work (Kroha and Baeza-Yates, 2005), we have been experimenting with statistical methods of text classification that are based on the frequency of terms to distinguish between positive news and negative news in terms of long-term market trends. In (Kroha and Reichel, 2007), we presented a grammar-driven text mining method, i.e. we have built a grammar that describes templates typical

for specific groups of news stories written in English. Each sentence in a news story is analyzed by a parser that determines the template to which the sentence belongs. Sentences and news are classified according to these assignments. We compared the statistical method and the grammar-driven method in (Kroha et al., 2007).

The first method (Kroha and Baeza-Yates, 2005), (Kroha et al., 2006) uses Bayes classification in two modifications. Its disadvantage is that it is based on frequency of words and does not respect the structure of sentences. The next method (Kroha and Reichel, 2007) is grammar-driven, i.e. it uses a grammar for description of classification so that it can mine more from sentences—so far from composed sentences. This method can be seen as a refinement of the first method.

3 GRAMMAR-DRIVEN METHOD FOR NEWS IN GERMAN

In methods of information retrieval, stop-list and term frequency are used to specify the content of documents. The goal is to find documents that concern the given topic which is given by some terms, too.

For the purpose of text classification, term frequency is not enough and some words from the stop-list can completely change the classification. Some features that are important for the classification are given by the sentence structure and not by the term frequency. In (Kroha and Baeza-Yates, 2005), we presented an example in which two news stories have the same term frequency but a completely different meaning.

To overcome the problem presented above, we have written a grammar describing grammatical constructions in English (Kroha and Baeza-Yates, 2005) that usually bring positive or negative meaning to a sentence.

3.1 Data for Grammar Construction

To get an overview what form a positive or a negative news story in German can have, we analysed by hand 547 market news having 8448 sentences. News and sentences are not only positive or negative. Many of them do not have clearly defined meaning, i.e. it cannot be clearly decided what an influence they can have. Because of that we filtered out all sentences that did not have a company name (no company news) as a subject, all sentences that only commented the movement of indices, all sentences having political content, all sentences containing speculations and guesswork.

This way we filtered out about 64 % of news stories. Some news use the company name only in the first sentence and the following sentences can be difficult to assign (roundup news). Some sentences describe events but it is difficult to decide in which way they influence the markets, e.g. company take-over, suspension of staff, paying a penalty. Some sentences can be clearly identified as positive or negative. They were only 514 of 8,448. Many sentences could not be assigned to any pattern and were marked as “not classified”. They were 2,240 of 8,448. The overview is given in Table 1.

Starting from the news and sentences analysed manually, we defined terms and structure of sentences that should be represented in a grammar.

Example:

```
Start = {System.out.println
        ("Classification: ");}
      (Sentence)* ;
Sentence = Positive | Negative;
Positive = {System.out.print
            ("test positive");}
          Company Subject PosVerb '.'
          {System.out.println
            (" > is positive");};
Negative = {System.out.print
            ("test negative");}
          Company Subject NegVerb '.'
          {System.out.println
            (" > is negative");};
Company = 'Allianz' | 'Bosch';
PosVerb = 'steigern' | 'steigert';
Verb = 'verzeichnen' | 'verzeichnet';
Subject = 'Gewinn' | 'Umsatzrueckgang';
```

3.2 Precision of our Grammar

The used grammar was build with messages from November 1999 and tested with messages from November 2004. To avoid a lot of sentences getting filtered by the withoutcompany-filter, a set of companies and company paraphrases were also inserted into our grammar. To check how accurately it works, we compared every grammar-found classification with the preceding sentences by hand. In fact, 89 % of all sentences were correctly matched by the grammar in our training period. As much as 91 % compliance could be achieved for the test period.

Considering the huge number of messages, it has been impossible to get a recall about sentences that should have been matched by the grammar but have not been.

Table 1: Types of news stories.

| Type | Number of News | Number of Sentences | Filtered out |
|--------------------------------|----------------|---------------------|--------------|
| Company news | 297 | 3878 | 53 % |
| Roundup news | 30 | 687 | 53 % |
| Overview | 59 | 1612 | 55 % |
| Currency news & Commodity news | 128 | 1831 | 91 % |
| Political news | 33 | 440 | 98 % |
| Σ | 547 | 8448 | 64 % |

4 IMPLEMENTATION

The described system was implemented as a component of the information system WEBIS in Java under Eclipse in (Nienhold, 2009). The BNF-form of the grammar was transformed into an executable version of a parser. We used the tool Bex (Franke, 2000) that produced the source code of the corresponding ll(k)-parsers. This source code was adapted by a classification object `ClassifyObj(c)` that completed the grammar by semantic rules describing the classification of messages.

The grammar can be easily created and edited using the grammar editor we wrote. It is possible to enhance new keywords and patterns to match more sentences. Via an import interface the opportunity is given to add several companies that are listed in a text file. In fact, the result will become more exact.

The lexical analysis has to transpose the textual messages into a form suitable for processing by the parser, i.e. into a form that will be accepted by all means (the case `notClassified` is a part of the grammar). It is necessary for every sentence to be classified because the parser should not interrupt the classification.

At the very beginning, each message will be decomposed into sentences. All abbreviations are investigated for change (to delete the dot) to guarantee that sentences are identified correctly. Then all words will be removed from each sentence that are not terminals of the given grammar. This process of text message preparation can be called normalization. In the next step normalized texts are processed.

The process of classification runs on normalized text messages. The parser reads every message and builds a corresponding object `ClassifyObj` that contains the number of classified (event. non-classified) sentences. This means that every classification of a sentence is stored (negative, positive or non-classified) and—as a result of a majority decision—the whole message gets a classifier, too. All details are described in (Nienhold, 2009).

5 EXPERIMENTS AND RESULTS

We used about 554,000 market news in German covering the interval from November 1999 to June 2009. It was not possible to recognize and classify all sentences completely. In the set we checked, we achieved a recall of about 73 %, i.e. 27 % of the relevant sentences were not matched by the grammar, and a precision of about 88 %, i.e. 12 % of all sentences were classified wrongly. The data processing was time consuming. In the first approach we used a PC having Intel Core 2 DUO E6750 processor with 2 GB memory. It took 21 hours. In the next approach we used an Apple MacPro computer having 8 kernels, we wrote our programs using threads. The process took 4 hours then.

The classification found was used to calculate a prediction. We weighted the positive news with +1, the negative news with -1, not classified news with 0. Then, we can use the following formula:

$$Prediction_t(pos, neg) = \frac{h_t(pos) - h_t(neg)}{h_t(pos) + h_t(neg)} \quad (1)$$

In the equation above (1) $h_t(pos)$ is the number of positive classified news at time point t and $h_t(neg)$ is the number of negative classified news at time point t .

The obtained prediction function (smoothed) in context of the german DAX market index is shown in Fig. 1. In region A (November 1999 - March 2001), we can see that the DAX and our prediction are practically identical. In region B (April 2002 - June 2003), we can see the breaking point of our prediction curve get ahead of the breaking point of DAX by about nine months. In region C (January 2007 - March 2009), we can see again that the breaking point of our prediction curve gets ahead of the breaking point of DAX by about a half a year, i.e. the minimum of the prediction curve comes at end of September 2008 and the minimum of DAX comes at the end of March 2009.

We have got a more precise prediction—shown in Fig. 2—when we calculated the prediction only from news that concern companies obtained in index DAX.

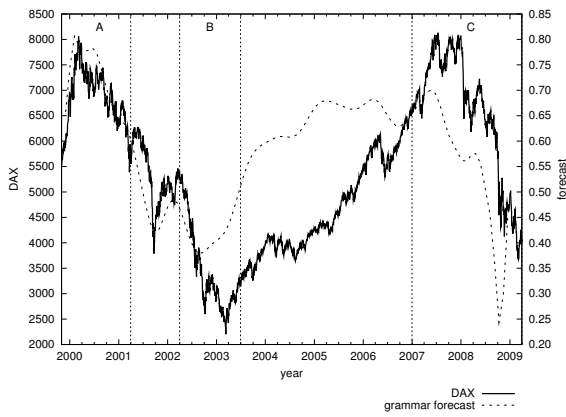


Figure 1: DAX and prediction smoothed - All market news.

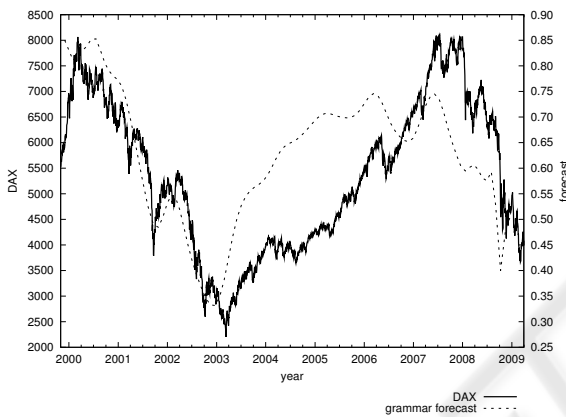


Figure 2: DAX and prediction smoothed - DAX market news.

In this case, our prediction get six month ahead of DAX in both cases B and C.

5.1 Statistical Analysis

In the previous section, we predicted two breaking points of type from-Down-to-Up with a forward of half a year to nine months. As we can see in Fig. 2, we are not able to predict the breaking points of type from-Up-to-Down.

An interesting topic is now to investigate to what extend it is possible to predict the DAX within n weeks in the future, i.e. what is the precision of the forecast and how it depends on time. We compared all existing trends of the DAX weekly with the trends of our prediction in a time intervall by [-20 weeks, +20 weeks]. The result is shown in Fig. 3.

The resulted pseudo code represents the way we have obtained values given in Fig. 3 :

```
for i=-20 to 20 step 1
  for n=firstWeek to lastWeek step 1
```

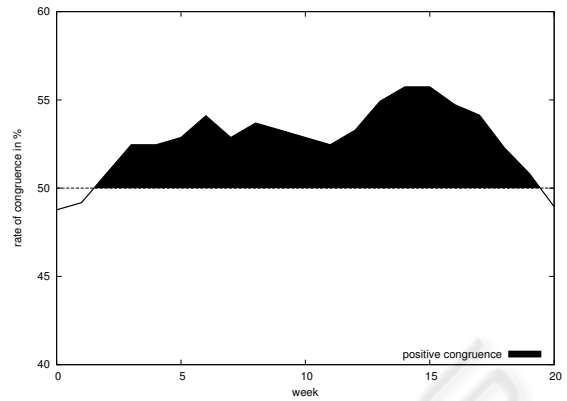


Figure 3: Congruence between the prediction and the DAX trend.

```
if(lastWeek < n+i < firstWeek)
  continue
if(DAXweek_n = uptrend)
  if(PREDweek_(n+i) = uptrend)
    equal++
rate_of_congruence_i = equal/n
```

Hence, we got an overcome by 55.7 % in +14 and +15 weeks. So we can say that if our prediction shows an uptrend in 14 weeks the DAX will have an uptrend by 55.7 %, too.

6 CONCLUSIONS

As we have shown, our prediction calculated from the quotient of good and bad news has interesting properties. In the last ten years, we have two breaking points of long-term trends changing from down to up. In both cases, our prediction curve obtained from news about companies in DAX gets six months ahead of the DAX index which is the main index of the market in Germany. Of course, two cases are not enough and we cannot state it as a rule. Additionally, there are other influences not contained in market news.

The reason why we can predict only from-Down-to-Up changes is very probably a psychological one. Frustrated investors observing markets going down do not want to believe that there are some positive changes in news and this is why the change of news goes ahead of market change.

In further work, we will try to evaluate the impact of messages in context of other economic parameters. We are sure that the same news story will be interpreted by investors differently when it comes in a different context of other economic parameters. We are just experimenting in including parameters concerned the classified news as inputs into a neuronal network.

In such a way, we can observe not only the influence of market news but also the influence of other economic parameters like oil price, gold price, relation between currencies, etc. We hope to get some refinement of the prediction.

The next problem is that we also need to take into account that some of news are not true and some of them has been constructed with the intention to mystify the investors. The task to filter out such news would be very interesting but too difficult at the very moment.

Also, we will try to evaluate a correlation between specific events of messages over a short and long timeperiod, i.e. fulfilled prognoses. An event weighting will get more precisely using weights in an interval of [-1, +1] instead of -1, 0 and +1.

REFERENCES

- Fama, E. (1970). *Efficient capital markets: A review of theory and empirical work*. Journal of Finance, 25, pp. 383-417,.
- Franke, S. (2000). Bex a bnf to java source compiler. In <http://www.bebbosoft.de/>.
- Kroha, P. and Baeza-Yates, R. (2005). A case study: News classification based on term frequency. In *Proceedings of 16th International Conference DEXA'2005, Workshop on Theory and Applications of Knowledge Management TAKMA'2005*, pp. 428-432. IEEE Computer Society.
- Kroha, P., Baeza-Yates, R., and Krellner, B. (2006). Text mining of business news for forecasting. In *Proceedings of 17th International Conference DEXA'2006, Workshop on Theory and Applications of Knowledge Management TAKMA'2006*, pp. 171-175. IEEE Computer Society.
- Kroha, P. and Reichel, T. (2007). Using grammars for text classification. In *Cardoso, J., Cordeiro, J., Filipe, J.(Eds.): Proceedings of the 9th International Conference on Enterprise Information Systems ICEIS'2007, Volume Artificial Intelligence and Decision Support Systems, INSTICC with ACM SIGMIS and AAAI*, pp. 259-264. INSTICC with ACM SIGMIS and AAAI.
- Kroha, P., Reichel, T., and Krellner, B. (2007). Text mining for indication of changes in long-term market trends. In *Tochtermann, K., Maurer, H. (Eds.): Proceedings of I-KNOW'07 7th International Conference on Knowledge Management as part of TRIPLE-I 2007, Journal of Universal Computer Science*, pp. 424-431.
- Lavrenko, S., Lawrie, O., and Jensen, A. (2000). Language models for financial news recommendation. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pp. 389-396.
- Macskassy, S. and Provost, F. (2001). Intelligent Information Triage. In *Proceedings of SIGIR'01, New Orleans, USA*.
- Nahm, M. (2002). Text mining with information extraction. In *AAAI 2002, Spring Symposium on Mining Answers from Texts and Knowledge Bases, Stanford*.
- Nahm, U. and Mooney, R. (2001). Mining Soft-Matching Rules from Textual Data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence IJCAI-01*.
- Nienhold, P. (2009). *Classification of News by a ll(k)-Grammar as a Component of WEBIS*. M.Sc. Thesis, Department of Computer Science, TU Chemnitz, Chemnitz, (in german) edition.
- Shleifer, A. (2000). *Inefficient Markets – An Introduction to Behavioral Finance*. Oxford University Press, Oxford.