Prediction of Patterns in Financial Time Series in Response to External Events

Bachelor Thesis

Alexei Pankov

Faculty of Economics
National Research University
Higher School of Economics

Supervisor: Alexey Artemov, Ph.D.,
Associate Professor, Faculty of Computer Science
Researcher, Skolkovo Institute of Science and Technology

This is a work-in-progress version of the thesis. Some pieces are subject to change. Please ensure you have the most recent edition:
`https://github.com/apnkv/hse/raw/master/thesis.pdf`

Contents

1 Introduction **4**

1.1 Background 4

1.2 Problem statement 6

2 Literature review **6**

3 Data preparation **9**

4 S&P 500 **11**

4.1 Model selection 11

4.1.1 Deterministic baseline 13

4.1.2 Non-trivial baseline 13

4.1.3 Models with textual features 16

4.2 Model evaluation 18

5 Moscow Exchange index **18**

6 Methodology **19**

6.1 Key definitions 19

6.2 8-K forms 20

6.3 Tools and algorithms 21

6.3.1 Random forest 21

6.3.2 Pointwise mutual information 21

6.3.3 TF-IDF transformation 22

6.3.4 Feature selection and dimensionality reduction . . 22

7 Conclusion **23**

References **24**

1 Introduction

1.1 Background

Stock markets

Stock markets are considered to represent current economic conditions and react lively to their changes. It is observed that stock prices respond immediately to key economic and political events as well as company reporting, whereas it can take weeks to months for them to be reflected in conventional measures of economic activity, such as macroeconomic statistics, business charts, investment ratings, etc.

For various agents, it is desirable to know how exactly various events affect financial markets. Such information can be used to gain extra profits in trading, to evade excess risk in portfolio management, to plan investment strategies against various hypothetical scenarios, to design optimal economic policies, and in many other applications.

Typical measurable stock market reaction to events include immediate price changes, periods of elevated returns or volatility, and others. There can be different reasons of such reaction. Many studies have emphasised the impact of macroeconomic news releases on stock market movements, a survey of such examples can be seen in (Funke and Matsuda 2006). Such news include changes in interest rates, publication of

official GDP growth estimates, unemployment numbers, and others.

Non-numeric information

Many different approaches have been devised to quantify inflows of non-numeric information to the market. For a particular company stocks, a rather simple approach is using the number of mentions of the company in the news. For example, Alanyali, Moat, and Preis 2013 show that the count of mentions of a company in the daily news is correlated with trading volumes of the corresponding stock on that day and the day that follows (and also suggest that a correlation with absolute returns may exist).

Most recent studies make use of the more advanced data mining and natural language processing (NLP) techniques. Algorithms purposed for sentiment analysis, topic modelling, named entity recognition etc. are used to extract structured data for problems of returns time series regression, categorical prediction (e. g. whether the stock price is going to rise/fall/remain steady), and econometric analysis involving hypothesis testing about distribution parameters (see Nassirtoussi et al. 2014 for a survey on text mining in market prediction).

1.2 Problem statement

This paper reduces the scope to a particular task of categorical prediction: forecast the direction of the price relatively to the current one, given that a company has recently published its earnings figures. The research will build upon a framework suggested by H. Lee et al. 2014. The hypothesis is that textual features will improve the predictability in this classification task.

The paper is set as follows. In Section 2, a literature review is presented. In Section 3, details are given on data preparation. In Section 4 an experimental setup from the original paper is reproduced, with detailed description of data retrieval and preprocessing. It is verified by the reproduction of original results. An evaluation of the model is performed using updated dataset versions. In Section 5, an attempt is made to apply the model to the Russian market, specifically, to the constituents of the MOEX index. Section 6 describes key definitions, methodological tools and algorithms that are used throughout the work.

2 Literature review

Explanation and prediction of stock markets behaviour has been a popular topic of economic analysis for a long time. Initially, it was widely

believed that financial time series follow a random walk pattern and thus cannot be predicted fundamentally. In 1965, the efficient market hypothesis was introduced by Eugene Fama, stating that current stock prices incorporate all information about previous events and change only in response to unexpected events as information becomes available to the market. The key point behind the efficient market hypothesis is that information is a very, if not the most, important component in prices. Though this statement implies that it is very difficult to outperform the market (because once someone has the relevant information, he or she immediately applies this knowledge to take advantage in the market), it does not rule out the theoretical possibility that relevant information might already be out there, concealed in immense streams of data that nobody has handled yet.

Correct use of existing and incoming information is discussed in the context of trading strategy planning (see Hong et al., 2000) and considered in hedging models, policy planning and many other cases (see Funke and Matsuda 2006).

Another direction of research of stock prices behaviour stems from the very process of investment and trading. Traders buy and sell in response to information they get from news regarding their assets of interest. Thus, in principle, the news should bear necessary information to predict market movement to some extent. This role of news has been emphasised many times in the literature, for instance, in (Pearce, 1984;

Poterba et al., 1988) in the context of returns and in (King, 1990) in the context of volatility. However, there is no complete understanding, let alone consensus in how to embed news into computations. Many different approaches have been devised to quantify information inflows to the market, including construction of numerical proxies such as amount of mentions of the company's name in daily news (Alanyali, Moat, and Preis 2013), trading volumes and others. While these news-derived indicators are helpful in discovering some useful correlations and some prediction tasks (for example, trading volumes are significant in volatility forecasting [Lamoureux & Lastrapes, 1990]), they lack consideration of probably the most important features of news: their semantic content and described relations between economic actors.

Advances in this direction were made with the emergence and development of natural language processing methods and tools. Many solutions have been introduced: for instance, representing documents as count distribution of encountered words (or a so-called bag of words) and extracting event information (Ding et al., 2016). One particular useful NLP approach is to create word- and document embeddings: their representations in vector spaces such that their proximity can be compared according to some metric corresponding to such space. This way, one can, for example, tag news as positive or negative, or know the topic of a piece of news: say, acquisition or expansion to a foreign market. For example, sentiment analysis has been widely used, yielding a stable

accuracy improvement in the task of forecasting immediate stock price direction (see Bollen et al., 2011). Such NLP-based predictive techniques have been discussed primarily within the problem of stock price movement direction in the context of English language and Western markets (see Nassirtoussi et al. 2014 for reference).

3 Data preparation

Several datasets of texts are used throughout the work for different cases. One is the collection of 8-K filings, provided by H. Lee et al. 2014. In this dataset, 8-K texts are annotated with timestamps of their publication and the form fulfillment reason, making it comfortable to align them with price changes and generate valuable categorical features. However, these texts needed to be preprocessed.

The texts were cleaned from tables and numeric values since these cannot be used efficiently without sophisticated knowledge mining systems. Also the texts contained remainders of the user-related interface strings unrelated to financially important content, e. g.:

Check the appropriate box below if ... any of the following provisions:

- Written communications pursuant to Rule 425 under the Securities Act (17 CFR 230.425)
- Soliciting material ... (17 CFR 240.14a-12) ...

Tables, numbers and unrelated content were removed using Python programming language. The total number of 8-K filings in the dataset is approximately 180 000 entries.

[TODO: add some descriptive statistics]

For Russian case, there also exist forms that are completed in case of significant corporate event. They are published continuously on (*e-disclosure.ru* 2018) along with timestamps and event categories. The number of such forms available for each company ranges from tens to thousands. Their texts are usually not as informative as typical 8-K filings, because frequently they consist of rather bureaucratic language amid a lot of insignificant information, and a large part of it is purely organisational. See an example:

```
1. Общие сведения 1.1. Полное фирменное наименование
эмитента (для некоммерческой организации -
наименование): Акционерное общество "ННК-Хабаровский
нефтеперерабатывающий завод" 1.2. Сокращенное
фирменное наименование эмитента: АО "ННК-Хабаровский
НПЗ" 1.3. Место нахождения эмитента: 680011,
Российская Федерация, г. Хабаровск, ул. Металлистов,
17. 1.4. ОГРН эмитента: 1022701129032 1.5. ИНН
эмитента: 2722010040 ...
```

Thus, event type features will be taken into the dataset, while 8K texts will be substituted with news occurred on or just before the event date.

Another text dataset used in the case of Moscow Echange is a collection of headlines relevant to the constituents of the MOEX index, obtained from Finam site, where they have been pre-selected to appear as at least somewhat relevant to the target stock.

4 S&P 500

4.1 Model selection

First, of all 8-K forms, only those corresponding to companies in S&P 500 were left. Data on prices and earnings was subsampled so that the

dataset is complete, i. e. each observation has a price record, an 8-K text and an EPS surprise figure.

The resulting dataset consists of 14 607 samples. The samples were divided into three classes as follows. First, the series of returns were calculated for each stock in the data. The returns were normalized to the S&P 500 index, as described in the methodology section. Depending on normalized returns, the following classes were assigned to each observation: UP if normalized returns $\tilde{r}_t > 0.01$, DOWN if $\tilde{r}_t < -0.01$ and STEADY otherwise. Such division yielded an approximately equal distribution of the classes, which can be seen in the following table.

	UP	STEADY	DOWN
count	5142	4674	4791
percentage	35.2 %	32.0 %	32.8 %

Table 1: Normalized returns class distribution

Further, the data were separated chronologically. Years 2011–2012 were put out for testing, 2009–2010 for hyperparameter tuning, and 2002–2008 were used as a training set. Among the subsets, class distribution is very similar to the general sample.

4.1.1 Deterministic baseline

To ensure the following experiments, a very simple yet powerful baseline was applied: predict UP if EPS surprise was positive, DOWN if negative and STEADY if the consensus earnings forecast was at par with actual called earnings. This baseline turned out to show accuracy around 47.2% on the testing set and around 44.3% on the whole sample. It is worth commenting on class recall—how many samples of a class were labelled as belonging to this class: A rather high recall of the UP class tells us that

	UP	STEADY	DOWN
Recall	78.4 %	13.9 %	29.8 %

Table 2: Class recall by pure-EPS baseline

when a firm exceeds its earnings estimations, it is often a positive market signal: in such situations, returns on its stocks tend to outperform the index. Another thing to be learned is from low recall on the DOWN class: unmet earnings expectations are not the only reason for a stock price to fall.

4.1.2 Non-trivial baseline

Several financial features were generated and collected prior to text processing to test another purely financial, but extended baseline, including cumulative (normalized) returns over a month, a quarter and a year,

8-K event category labels and returns on the VIX ticker, a volatility expectation index.

Feature selection was performed on 8-K fulfillment reason categories, since not all of them are connected with investors' expectations and thus not all of them are relevant in price direction forecasts. It was done by calculating so-called Pointwise Mutual Information for each of the category features with the class vector.

This measure reflects codependency between samples of two random variables. Thus, choosing categories with the greatest PMI, we filter our feature set to contain features more connected to the target variable. Using this approach, fifteen most relevant categories were chosen (see table 3).

Out of all event category columns, only these were kept. Eventually, the dataset for the second model contained the EPS surprise column, 15 category dummy columns, 3 columns of cumulative returns, and the returns on the VIX index.

To address the classification task in a non-trivial case, several algorithms were assessed, including a logistic regression, decision trees, SVM and random forests. The highest accuracy on the hyperparameter tuning set was achieved by a random forest with 2050 individual tree estimators, assigning not fewer than 15 samples to each leaf node, making splits of not less than 250 samples on each step. A detailed explanation of these

Regulation FD Disclosure
Material Impairments
Cost Associated with Exit or Disposal Activities
Entry into a Material Definitive Agreement
Termination of a Material Definitive Agreement
Non-Reliance on Previously Issued Financial Statements or a Related Audit Report or Completed Interim Review
Material Modifications to Rights of Security Holders
Changes in Registrant's Certifying Accountant
Submission of Matters to a Vote of Security Holders
Unregistered Sales of Equity Securities
Change in fiscal year
Amendments to Articles of Incorporation or Bylaws
Temporary Suspension of Trading Under Registrant's Employee Benefit Plans
Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet Arrangement of a Registrant
Results of Operations and Financial Condition

Table 3: Most relevant 8-K categories

parameters is given in the methodology section. On the test set, the top-performing random forest classifier achieved accuracy of 49.9 %.

[TODO: Paste classifier selection tables and graphs]

[TODO: Provide an example of a tree traversal]

4.1.3 Models with textual features

To employ text features from 8-K forms, a text processing pipeline was constructed. Firstly, the texts were cleaned from metadata, timestamps and unrelated service pieces as described in the previous section. All words were cut down to their stem so as various forms of one word as one term, for instance, 'acquired' $\rightarrow$ 'acquir', 'acquiring' $\rightarrow$ 'acquir'. All words that encountered in all texts fewer than ten times were removed, leaving around 5000 features, which obviously required factorization. All auxiliary and frequently occurring words with little meaning were removed according to the following list:

```
stopwords = (‘the’, ‘to’, ‘of’, ‘by’, ‘on’, ‘this’,
‘be’, ‘it’, ‘report’, ‘and’, ‘as’, ‘in’, ‘a’, ‘has’,
‘with’, ‘or’, ‘from’, ‘that’, ‘are’, ‘at’, ‘not’,
‘will’, ‘an’, ‘were’, ‘more’, ‘herein’, ‘hereunto’,
‘all’, ‘over’, ‘no’, ‘hereto’, ‘do’, ‘exchang’,
‘secur’, ‘act’, ‘for’, ‘is’, ‘behalf’, ‘registr’,
‘financi’, ‘statement’, ‘may’, ‘was’, ‘million’,
‘per’, ‘our’, ‘year’, ‘these’, ‘requir’, ‘caus’,
‘pursuant’, ‘includ’, ‘result’, ‘compani’, ‘author’,
‘sign’, ‘undersign’, ‘duli’, ‘quarter’, ‘oper’,
‘other’, ‘file’, ‘exhibit’, ‘inform’, ‘releas’,
‘share’, ‘which’, ‘materi’, ‘end’, ‘forward-look’,
‘we’, ‘busi’, ‘relat’, ‘use’, ‘billion’, ‘if’, ‘two’,
‘under’, ‘should’, ‘can’, ‘but’, ‘both’, ‘also’,
‘such’, ‘into’, ‘same’, ‘their’, ‘said’, ‘through’,
‘have’, ‘those’, ‘than’, ‘then’)
```

After that, for each text, a vector was constructed in a way that each coordinate corresponded to a unique term among all encountered in texts, and value in that coordinate was the number of occurrences of this unique term in the text.

[TODO: Describe TF-IDF transformation and NMF factorization and write more here]

After all text transformations, 50 features were obtained. After at-

taching them to the data and training a new classifier, an accuracy of 53,7 % was achieved.

The model with textual features turned out to be significantly better than both baselines. The statistical significance was evaluated using approximate randomization (Hoeffding 1952; Noreen 1989).

[TODO: Explain the testing method in detail]

4.2 Model evaluation

[TODO: Finish and describe an extension to 2002–2018]

5 Moscow Exchange index

In an attempt to conduct a similar research in the context of Russian markets, a smaller dataset was collected in the fashion of the original S&P 500 analysis. For the companies which constitute the MOEX index, consensus EPS forecasts were obtained using Bloomberg terminal. In the Russian case, this feature is very sparse, yielding only 366 samples for 30 firms.

In case if Bloomberg estimates did not provide a quarterly EPS consensus, the yearly one was used, divided by 4.

[TODO: Write this section]

6 Methodology

This research will be built upon H. Lee et al. 2014. In the paper, authors predict the direction of price movement after an earnings call using several financially motivated features along with text features extracted from 8-K reports, which are filled out by US firms on significant corporate events.

6.1 Key definitions

First, we will define some key financial terms. Let P_t be the price of an asset at time t. In the case of stocks, let P_t be the closing price. Then, *returns* on this asset are defined as $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$. Let us call *normalized returns* the value $\tilde{r}_t = r_t^{\text{asset}} - r_t^{\text{index}}$.

Earnings per share, or EPS, can be defined as the difference between net income and dividends on preferred stock divided by the average number of outstanding shares during the period.

Earnings surprise is the difference between earnings called by the company and mean consensus forecast of this value. Higher earnings surprises are believed to bring optimistic news, as it means that a com-

pany performed better than expected.

[TODO: describe VIX]

6.2 8-K forms

An 8-K form is a means of reporting which US companies are obliged to fill out and submit to the Security Exchange Commission on current events that can be significant to stakeholders. Table 4 presents some of such event types. Besides obligatory categorical description, 8-K forms

completions of the acquisition of assets
changes in financial conditions
results of operation
change in control board
election or department of directors
change in fiscal year
SEC investigations and checks
equity sales
modifications to shareholder rights
material direct financial obligations
change in accountant

Table 4: 8-K fulfillment reasons

usually include a written explanation and interpretation of the facts, for example:

```
IHOP Corp.  (NYSE: IHP) today announced a new strategic
growth plan designed to increase long term shareholder
value.  In 2003, IHOP plans to transition from
company-financed development of new restaurants to
a traditional franchise development model.
```

Thanks to the rather detailed structure of such reports, they can be used as a source of linguistic features for prediction.

6.3 Tools and algorithms

6.3.1 Random forest

Random forest is a learning algorithm which implements the voting of different decision trees trained on random subsets of features and samples. Samples can be bootstrapped to ensure randomness. The voting mechanism here is known as Breiman bagging (see Breiman 2001) and helps to smooth out errors of individual decision trees and yield a sensible prediction.

[TODO: Provide details on inner workings and parameters]

6.3.2 Pointwise mutual information

[TODO: Provide a description]

6.3.3 TF-IDF transformation

6.3.4 Feature selection and dimensionality reduction

Using features bound to individual terms as predictors can arise different problems. Words can be ambiguous in different context and carry little meaning when used without their usual collocations. Moreover, keeping too many features in the model can lead to the curse of dimensionality.

To achieve a more stable model, vectors obtained from texts are mapped to a smaller dimension. This is done using the technique called *non-negative matrix factorization.*

[TODO: cite an algorithm to find an NMF]

Non-negative matrix factorization is a feature extraction approach rather similar to PCA. It maps vectors to vectors of smaller dimensions via linear combinations, but all coefficients are kept non-negative.

It is believed that the non-negativity restriction is the reason why NMF performs better than PCA in tasks related to language and image processing.

Linear combinations of such features from words can be thought of as 'topics'—frequently co-occurring word clusters, and we usually model such topics as additive. For example, we would say `'appointment of directors'` is rather `'appointment' + 'director'` than `'appointment'`

`+ 'staff' - 'worker'`. The mapping achieved in such a way is smoother than that of a PCA. More details and some applications are described in D. D. Lee and Seung 2001.

7 Conclusion

[TODO: state that textual features are valuable and powerful, insert comment on the poorer performance on moex, limitations]

References

Alanyali, Merve, Helen Susannah Moat, and Tobias Preis (2013). "Quantifying the relationship between financial news and the stock market". In: *Scientific reports* 3, p. 3578.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

e-disclosure.ru (2018). URL: `http://e-disclosure.ru` (visited on 2018).

Funke, Norbert and Akimi Matsuda (2006). "Macroeconomic news and stock returns in the United States and Germany". In: *German Economic Review* 7.2, pp. 189–210.

Hoeffding, Wassily (1952). "The large-sample power of tests based on permutations of observations". In: *The Annals of Mathematical Statistics*, pp. 169–192.

Lee, Daniel D and H Sebastian Seung (2001). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems*, pp. 556–562.

Lee, Heeyoung et al. (2014). "On the Importance of Text Analysis for Stock Price Prediction." In: *LREC*, pp. 1170–1175.

Nassirtoussi, Arman Khadjeh et al. (2014). "Text mining for market prediction: A systematic review". In: *Expert Systems with Applications* 41.16, pp. 7653–7670.

Noreen, Eric W (1989). *Computer-intensive methods for testing hypotheses*. Wiley New York.