Towards a Benchmarking Framework for Financial Text Mining

Caslav Bozic[1], Ryan Riordan[2], Detlef Seese[1], and Christof Weinhardt[2]

[1] Institute of Applied Informatics and Formal Description Methods, KIT
{bozic, detlef.seese}@kit.edu

[2] Institute of Information Systems and Management, KIT
{ryan.riordan, weinhardt}@kit.edu

Summary. Different data mining methods for financial texts and various sentiment measures are described in the existing literature, without common benchmarks for comparing these approaches. The framework proposed in this paper and the corresponding implemented system facilitate combining more sources of financial data into comprehensive integral dataset. The use of the dataset is then illustrated by analyzing the candidate measures by estimating parameters of regression on different returns and other financial indicators that can be defined using system's novel data transformation approach.

1 Introduction

Different data mining methods for financial texts are described in the literature, and systems that apply these particular methods have been implemented. For an evaluation of the proposed methods authors use different approaches, some of them incorporating assumptions that do not conform to real financial markets' conditions. This paper presents the current status of a project attempting to offer a possibility to compare performance of these systems by offering a framework, a base dataset, and an implementation of the system according to design science guidelines from Hevner et al. (2004).

The research is a part of the project FINDS (Financial News and Data Services) initiated within Information Management and Market Engineering Graduate School at the Karlsruhe Institute of Technology. Project FINDS has the goal to conduct innovative research on the analysis of quantitative and qualitative information related to financial markets and to provide services that can help both researchers in financial field, and also professional traders (Bozic, 2009).

The paper is organized as follows: in section 2 we give an overview of the existing literature in the field. Section 3 describes data sources for the comprehensive research set, preprocessing methods used on data and gets the reader acquainted with the dataset by presenting some descriptive statistics. Section 4 illustrates the usage of the framework and the system by comparing four sentiment measures. Section 5 concludes the paper and describes the ideas for future research.

2 Related Work

Many researchers have studied how published news influence market reactions. The methodologies used range from a fairly simple content analysis using classical statistical tools, to complex machine-learning methods. The approaches vary from an engineering approach which focuses on implementation and proving economic relevance, to chiefly theoretical approaches whose goal is to describe underlying economic phenomena.

Mittermayer and Knolmayer (2006a) compare eight text-mining systems, including their own. Since more technical performance criteria are often missing, it is not possible to draw clear conclusions about relative performance. Wüthrich et al. (1998) classify news articles published overnight on web portals into three categories depending on their influence on the one of five equity indices: Dow Jones, Nikkei, FTSE, Hang Seng, and Straits Times. With this system they attempt to forecast the trend of the index daily value one day ahead. They use Naïve Bayes, Nearest Neighbour and Neural Net classifier, and a hand-crafted underlying dictionary. Lavrenko et al. (2000) use Naïve Bayes classifier to classify news articles from Yahoo!Finance into five groups, according to the influence on particular U.S. stocks. The features were determined automatically and the forecast horizon was from five to ten hours. Gidófalvi and Elkan (2003) use again naïve Bayes classifier with three categories to recognize articles which have bigger positive or negative influence on constituents of Dow Jones index. With features defined using mutual information measure they work on ten minutes aggregated intraday data. Fung, Yu, and Lam (2003) partially use commercially available text mining systems to predict a price trend for intraday market movements of some of the stocks listed on the Hong Kong Stock Exchange. For classification purposes they use support vector machines. Finally, Mittermayer and Knolmayer (2006b) propose a high frequency forecast system that classifies press releases of publicly traded companies in the U.S. using a dictionary that combines automatically selected features and a hand-crafted thesaurus. For classification the authors use the polynomial version of SVM.

While most of the works focus on predicting price trends of single stock or index, there are works that aim at determining influence of news releases to volatility. Thomas (2003) improves risk-return profile by exiting the market in case of news that are predicting high volatility, while Schulz, Spiliopoulou, and Winkler (2003) attempt to classify press releases of German public companies according their influence on volatility of stock prices.

Another group of publications not included in the survey by Mittermayer and Knolmayer (2006a) contains works that do not primary attempt to prove economical relevance of published text by evaluating specifically tailored trading strategies, but rather to find statistically relevant relations between financial indicators and sentiment extracted from the text.

Antweiler and Frank (2004) use Naïve Bayes and SVM classifiers to classify messages posted to Yahoo!Finance and Raging Bull and determine their sentiment.

They do not find statistically significant correlation with stock prices, but they find sentiment and volume of messages significantly correlated to trade volumes and volatility. In their methodological paper Das and Chen (2007) offer a variety of classifiers, as well as composed sentiment measure as a result of voting among classifiers. In the illustrative example they analyze Yahoo stock boards and stock prices of 8 technology companies, but they do not find clear evidence that the sentiment index can be predictive for stock prices.

There are two pivotal articles published in the Journal of Finance. Tetlock (2007) observes Wall Street Journal's column "Abreast of the Market", uses content analysis software General Inquirer together with Principal Component Analysis approach and finds that high pessimism in published media predicts downward pressure on market prices. Authors of Tetlock, Saar-Tsechansky, and Macskassy (2008) succeeded to find that rate of negative words in news stories about certain company predicts low earnings of the company.

If we observe text mining methodologies as a transformations that assign numerical value to every textual string, we can refer to that numerical value as sentiment index. All publications from the former group have at least implicit statements about the predictive power of the specific sentiment index on e.g. returns or volatility. Following the evaluation approach from the latter group of publication, we aim at providing financial text mining research community with a framework and a tool that can be used for proving their statements using statistical significance criteria.

3 Data

3.1 Data Sources

We use Thomson Reuters TickHistory data and the output from the Reuters NewsScope Sentiment Engine as a source dataset. These data sources are convenient because they provide access to trading data over period of more than 10 years, and extensive amount of sentiment data related to financial news stories.

Data constituting the output of the Reuters NewsScope Sentiment Engine represents the author's sentiment measure for every English-language news item published via NewsScope in years 2003-2008. The measure classifies a news item into one of three categories: positive, negative, or neutral. The probability of the news item falling into each of the categories is also given.

Thomson Reuters TickHistory data is available through the DataScope platform. We use daily data for the period equivalent to the NewsScope Sentiment Engine dataset. This provides us with data on opening and closing prices for the particular product, bid and ask, as well as volume data. Other types of data about companies coming from this source are the total amount of shares, used for calculating market capitalization, and paid dividends, that can be used for adjusting the returns. Additionally, it is the source of the data about daily values of the MSCI Indices for individual countries, as well as MCSI World Index.

Field Name	Description	Source
ric	Reuters Instrument Code	TH
d	Date	TH
Open	Opening daily price	TH
High	Maximal price within the day	TH
Low	Minimal price within the day	TH
vol	Daily trading volume	TH
Last	Closing daily price	TH
Bid	Average bid	TH
Ask	Average ask	TH
spread	Average spread	TH
cc	Close to close daily return of the equity	D
oc	Open to close daily return of the equity	D
oo	Open to open daily return of the equity	D
co	Close to open daily return of the equity	D
cnt	Number of news items mentioning company	RNSE
pnaccnt	Number of news stories mentioning company	RNSE
nsent_pos	Number of positive news items mentioning company	RNSE
nsent_neut	Number of neutral news items mentioning company	RNSE
nsent_neg	Number of negative news items mentioning company	RNSE
avgsent	Average sentiment	RNSE
asent_pos	Average probability that news items are positive	RNSE
asent_neut	Average probability that news items are neutral	RNSE
asent_neg	Average probability that news items are negative	RNSE
net_sent	Net sentiment: asent_pos - asent_neg	D
net_sent_std	Standard deviation of net sentiment	D
ric_market	Abbreviation of company's home market	TH
country	Country	D
icc	Close to close daily return of the country index	D
ioc	Open to close daily return of the country index	D
ioo	Open to open daily return of the country index	D
ico	Close to open daily return of the country index	D
cnt_country	Number of news items related to country	RNSE
avgsent_item	Average sentiment (country level)	RNSE
eccc	Excess daily close to close return w/r to country index	D
ecoc	Excess daily open to close return w/r to country index	D
ecco	Excess daily close to open return w/r to country index	D
ecoo	Excess daily open to open return w/r to country index	D

Table 1: Main fields and sources (TH - TickHistory, RNSE - Sentiment Engine, D - derived)

Reuters NewsScope Sentiment Engine data has two main properties - the timestamp of the news item publishing, and the related company mentioned in the news item. It is preprocessed in a way that the records are aggregated on the level of each company, and also calendar day - according to local time in force at the location of company's home market. In that way we get an average sentiment for a company for each day in 6 years period. The sentiments are expressed by different calculated values. The first aggregated sentiment measure is average sentiment class - where the classes are represented by values 1, 0, and -1 for positive, neutral, and negative class, respectively. Further measures are average probabilities that each news item falls into positive, negative or neutral class. As the Reuters sentiment index used for the evaluation later we adopted a deducted value defined as average probability of positive class minus average probability of negative class within one day.

From the data on daily prices we derive data on different returns: namely close-to-open, close-to-close, open-to-open, and open-to close returns. This is done considering only days with trades for particular equity or index. The generic way of defining calculated fields derived from source data is described in later section. The calculated returns are adjusted for paid dividends by increasing the price of share on the trading day following the dividend payment by the amount of dividends paid per share.

3.2 Descriptive Statistics

Reuters NewsScope Sentiment Engine data consists of author's sentiment measures of English-language news items published through Reuters NewsScope in the period from 2003 to 2006 inclusive. Each record represents a unique mention of the specific company, with a possibility of one news item relating to more than one company. In our dataset there are 6,127,190 records about 10,665 different companies. Figure 1 shows the top twenty companies with greatest number of records related to that company. According to Reuters news production process, several news items can make one single news story, e.g. short alert item is published immediately, and after some time extension of the same story is published as a new item, or in the case of corrections. In the available data average number of news items per story is 1.995, saying that in average two items make one news story.

Increase in data volume over the years is obvious and yearly record number doubles over the period of 6 years, as shown on Figure 2. Besides, on the same figure, one can follow increase in the partial volume of records related to different home markets. The information about a company's home market is an integral part of the Reuters code uniquely identifying each instrument. Figure 2 separately shows data about the ten markets with greatest changes over the years. It can be seen that the fraction of the two biggest US markets, NYSE and NASDAQ, in the total data volume has grown from something over 40% to little less than 60% in this 6 years period.

Each news item can have multiple tags called topic codes, and topic codes are grouped into categories. One of the categories represents countries, and can be used

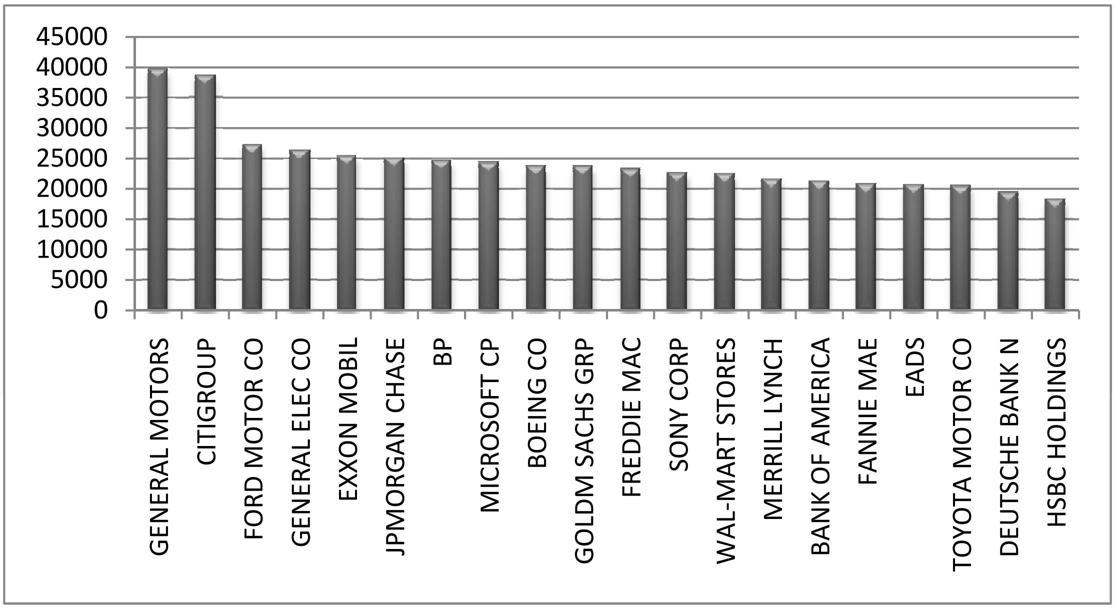

Figure 1: Top twenty companies by number of records

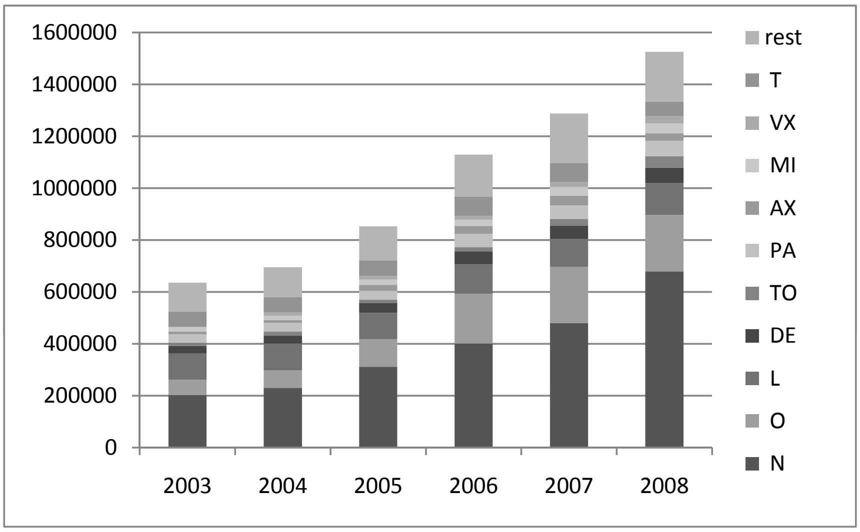

Figure 2: Data volume change over years

for determining what country is mentioned in the particular news item. Using this author tagging feature, average sentiment per country can be calculated. The Figure 3 is showing counties with greatest and lowest average sentiment, when country tags with less than 1000 mentions are excluded.

During the week, only 2.19% of all news items are published on weekends. Figure 4 shows the distribution of records per days of week. Our dataset comprises of data spanning 2192 days in total, considering GMT time zone, and there are even 13 companies that are mentioned at least once in more than 80% of all days, as shown in Table 2. One more interesting property of the data is noticeable sentiment decrease for news items published on Saturdays, as shown on Figure 5.

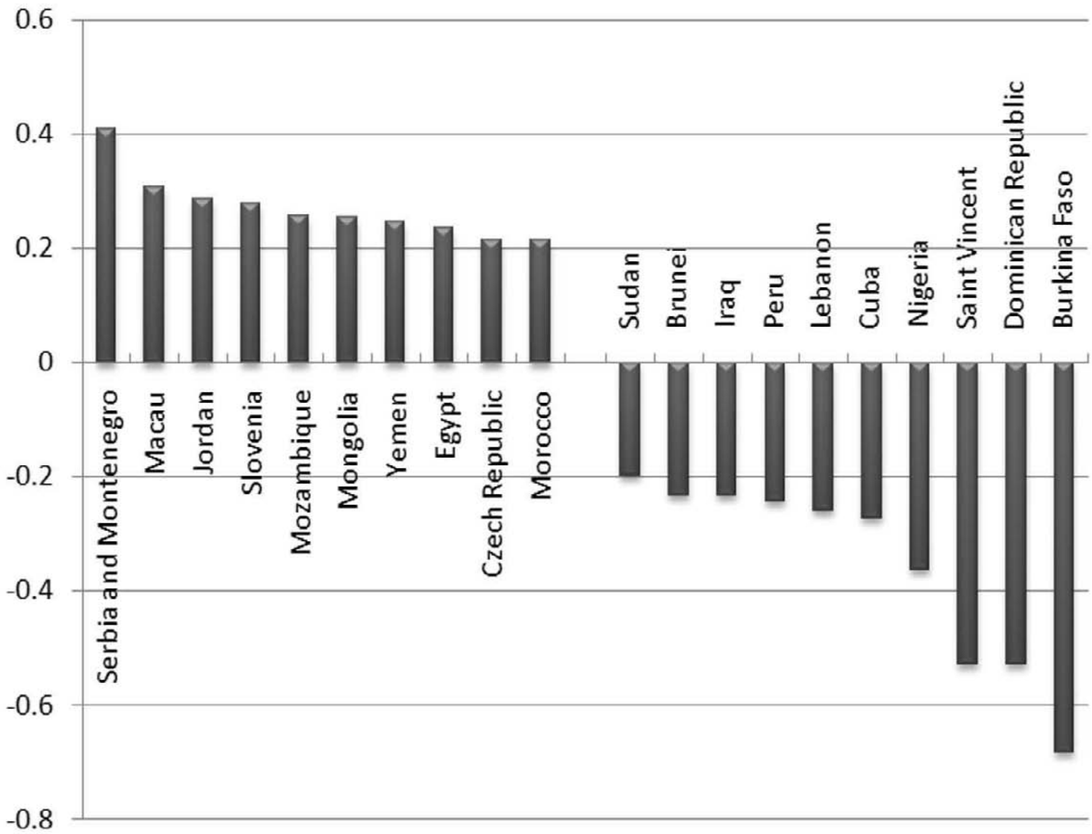

Figure 3: Best and worst average sentiment for countries with over 1000 mentions

Company	Days with news	Percent
GENERAL ELEC CO	1911	87.18
SONY CORP	1891	86.27
CITIGROUP	1863	84.99
BP	1851	84.44
WALT DISNEY CO	1841	83.99
GENERAL MOTORS	1837	83.8
EXXON MOBIL	1822	83.12
WAL-MART STORES	1802	82.21
BOEING CO	1798	82.03
MICROSOFT CP	1792	81.75
HSBC HOLDINGS	1791	81.71
GOLDM SACHS GRP	1782	81.3
FORD MOTOR CO	1754	80.02
EADS	1752	79.93
TOYOTA MOTOR CO	1729	78.88
DEUTSCHE BANK N	1729	78.88
VOLKSWAGEN AG	1720	78.47
VODAFONE GROUP	1711	78.06
CHEVRON	1704	77.74
TOTAL FINA	1693	77.24

Table 2: The 20 companies mentioned on most days

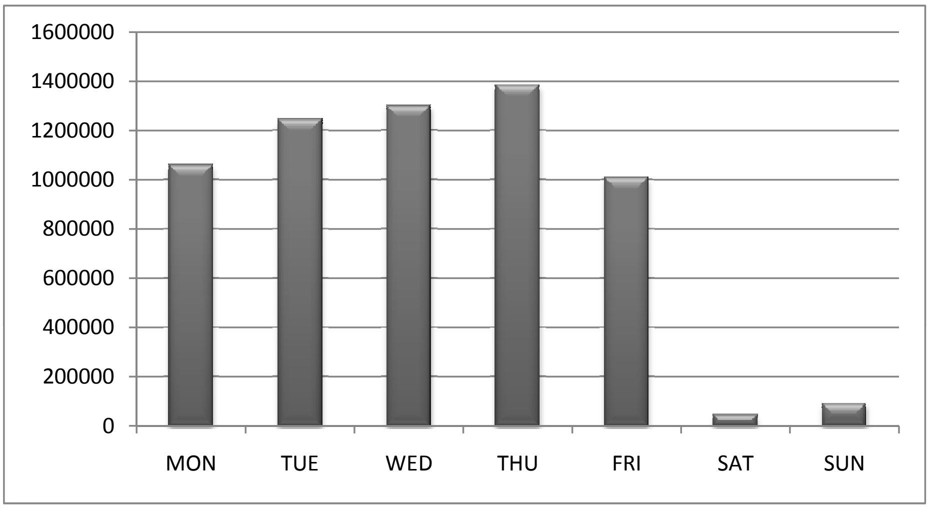

Figure 4: Number of records per days of week

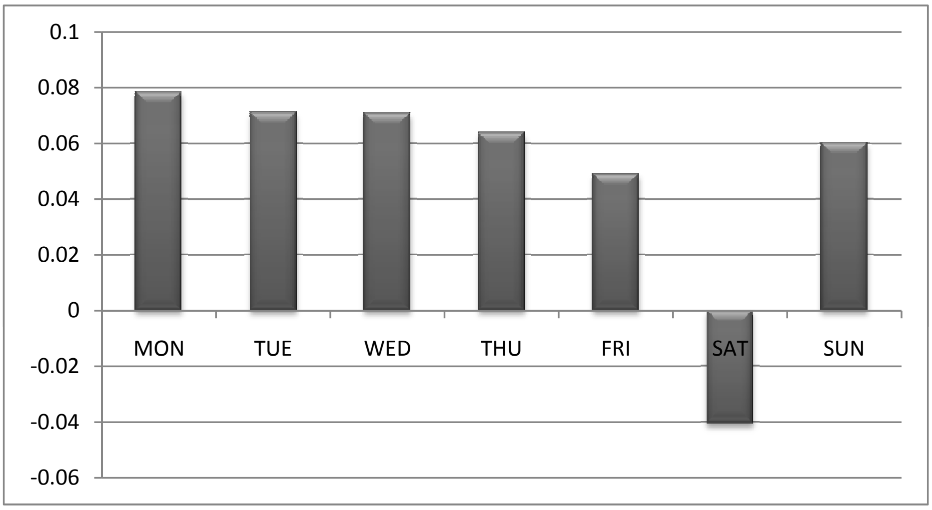

Figure 5: Average sentiment per days of week

3.3 Preprocessing

To enable easy extending of the existing dataset, generic interfacing is one of the system's features. It is generic in that aspect that it allows any format of the DataScope platform output to be loaded into the system automatically. New source to be added is placed in a comma separated values (CSV) file. The procedure of generic load then starts, making new data source available for the next steps of preprocessing. The names of the new columns that can be later used for referencing are extracted from header row of the comma separated input file. The necessary type transformations on the added data are conducted by means of metadata dictionary, which holds datatypes and formats related to the column names. In this way new data is transformed into the representation that can be used as input of the successive steps.

As our goal is to explore dependencies of various indicators and sentiment measures, another feature of the system is a novel approach to defining calculated indicators. Any of the columns from the source datasets can be used as a base for calculations. Besides full range of arithmetic operators, lagging operators can be defined and used as well. The definition of the calculated fields is given in the notation of simple grammar we defined. The definition is saved in control file, which is parsed and used for code generation. It produces PL/SQL code that applied to the source datasets produces additional datasets containing calculated fields.

Final result is a clean dataset with derived fields that were not part of the source datasets, designed for particular benchmarking purpose. The overview of the data flow can be seen on Figure 6.

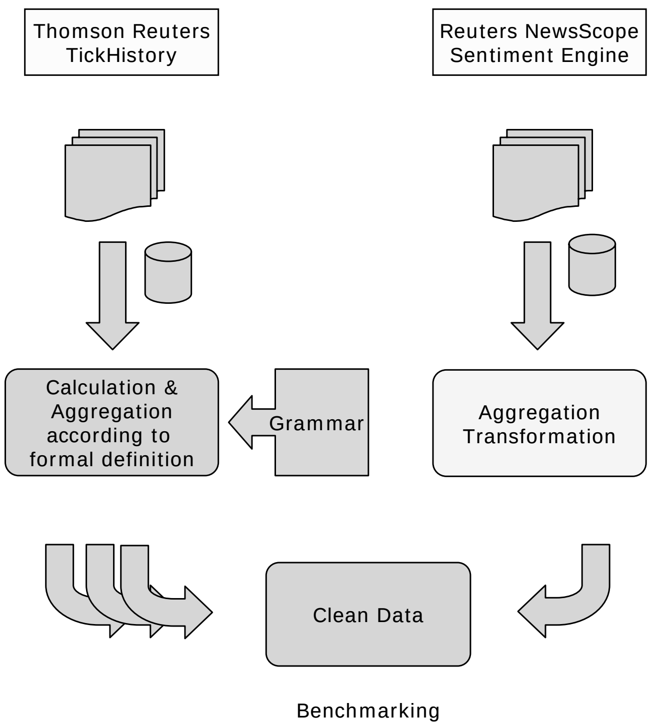

Figure 6: Overview of the Preprocessing system

4 First Results

As an illustration of framework application we performed comparison of four sentiment measures. Three of them are produced using classifiers that are implemented in the scope of FINDS Project. The first one is using Bayes-Fisher filter, second Support Vector Machines, while third implements artificial Neural Networks methodology. The classifiers are trained on high frequency intraday data on trades, aggregated to 1 minute average prices. The news items are grouped into classes depending on their influence on the price movements. Observing the period of two hours immediately following news item publication, we determine class of the news item. If the price stabilized on the level higher than the level at the moment of news item publication, news item was classified as positive. If the price stabilized on lower level, news item was classified as negative. If the price stabilization couldn't be observed, news item is considered neutral. More details about classifier implementation can be found in Pfrommer et al. (2010). Training period was first nine months of 2003, and news stories about five big technology companies (SAP, Oracle, Microsoft, IBM, and Apple) were used.

The last three months of 2003 represent the evaluation period, with total of 729 news stories analyzed. The process of news transformation shown on Figure 7 consists of text tokenizing, stemming, and finally text classification using selected methodology. Using the dataset and the framework we calculated parameters for regression of lagged dependent variables against each sentiment measure. The variables were four different daily returns, four daily excess returns, and average daily spread, all lagged up to ten days. The regression formula is given by $id_t = \alpha_{(id,i)} + \beta_{(id,i)} S_{t-i} + \varepsilon_t$ where id represents dependent variable, and i size of the lag. The results are presented in Table 3 and Table 4. The column title determines dependent variable, and the row represents days of lag between sentiment measure S and dependent variable id. Each cell shows two values: estimated coefficient value β first, and then p value according to t-statistics.

The results show that just for 14 variables could be proven statistically relevant dependence on lagged sentiment values. Although RNSE data shows most of the statistically significant results, seven, and all of them have positive coefficients, it is not possible to give clear statement about comparative performance of sentiment measures. This issue possibly arises from the low data volume, causing the statistical dependence not to be observed. To overcome the problem, regression parameter estimation for one of the measures is repeated with greater number of data points. RNSE measure is chosen because of the extensive volume of the data. Table 5 presents the results from regression parameters estimation on data sample including all six years of processed news stories in period 2003 to 2008 for two major U.S. markets contained in the system. All tested dependent variables show statistically significant dependence on sentiment value with lag of at least one day. The increase in number of statistically relevant results shows that greater amount of data offers a possibility to draw more confident conclusions.

Bayes-Fisher

lag	cc		oo		co		oc		spread		eccc		ecoo		ecco		ecoc	
1	-0.00824	0.11	-0.00316	0.54	-0.00641	0.15	-0.00182	0.48	0.00013175	0.45	-0.00646	0.23	-0.00439	0.40	-0.00386	0.43	-0.0026	0.24
2	-0.00212	0.72	-0.00557	0.19	-0.00046567	0.93	-0.00165	0.54	-0.00019337	0.18	-0.0039	0.53	-0.00263	0.47	-0.00252	0.65	-0.00147	0.54
3	-0.00056663	0.91	-0.00466	0.29	-0.00243	0.54	0.00187	0.49	-0.000633	0.01	-0.0028	0.53	-0.00096592	0.79	-0.00301	0.44	0.00029881	0.90
4	-0.00646	0.15	0.0026	0.51	-0.00667	0.10	0.00020725	0.93	-0.0004977	0.07	-0.00357	0.40	0.0008996	0.77	-0.0033	0.40	-0.00020749	0.91
5	-0.00032704	0.94	-0.00013232	0.97	0.00193	0.61	-0.00225	0.32	-0.00019928	0.30	0.00223	0.60	-0.00148	0.64	0.00242	0.51	-0.00016307	0.93
6	0.0034	0.42	-0.00163	0.67	0.0026	0.43	0.00079763	0.73	-0.0005966	0.04	0.00036264	0.92	-0.00081241	0.83	0.00037901	0.89	-0.00000455	1.00
7	0.00139	0.75	0.00642	0.12	-0.00234	0.47	0.00373	0.17	-0.00034695	0.26	0.00326	0.41	0.00584	0.17	-0.00092262	0.73	0.00424	0.11
8	0.00172	0.69	-0.00095658	0.81	0.00291	0.36	-0.00119	0.65	-0.000121	0.54	0.00165	0.68	0.00262	0.50	0.00106	0.70	0.0006209	0.80
9	-0.00343	0.27	-0.0057	0.11	0.00023065	0.93	-0.00366	0.08	-0.0004424	0.01	-0.00181	0.53	-0.00642	0.09	0.00206	0.42	-0.00385	0.05
10	-0.00411	0.19	-0.0038	0.25	-0.00204	0.49	-0.00207	0.21	-0.00007904	0.64	-0.00231	0.41	-0.00619	0.05	-0.00010074	0.97	-0.00218	0.12

SVM

lag	cc		oo		co		oc		spread		eccc		ecoo		ecco		ecoc	
1	0.00102	0.83	-0.00339	0.48	-0.00101	0.81	0.00203	0.40	-0.00003017	0.76	-0.00104	0.85	0.00127	0.81	-0.0027	0.59	0.00169	0.46
2	-0.0033	0.56	-0.00377	0.36	-0.00499	0.31	0.00169	0.52	-0.00004376	0.65	-0.00535	0.38	0.00371	0.30	-0.00637	0.24	0.00105	0.65
3	-0.00325	0.48	-0.00313	0.46	-0.00568	0.14	0.00243	0.36	-0.0000986	0.59	-0.00353	0.43	-0.00149	0.68	-0.00478	0.21	0.00128	0.58
4	-0.00281	0.54	-0.00186	0.64	-0.0045	0.28	0.00169	0.45	-0.00004667	0.83	-0.00236	0.58	-0.00109	0.73	-0.00413	0.30	0.00179	0.34
5	-0.0002228	0.96	-0.00054979	0.89	-0.00288	0.46	0.00266	0.25	0.00000991	0.96	0.00027364	0.95	0.00119	0.72	-0.00215	0.56	0.00239	0.22
6	0.00218	0.62	0.00156	0.69	-0.00253	0.47	0.00471	0.05	-0.00026022	0.41	0.00148	0.72	0.0039	0.35	-0.00121	0.70	0.0027	0.28
7	-0.00014802	0.97	0.0021	0.63	-0.00358	0.29	0.00344	0.22	-0.00000721	0.99	0.0029	0.48	0.00492	0.26	-0.00128	0.66	0.00416	0.13
8	-0.00127	0.78	-0.00473	0.25	-0.00147	0.66	0.00019947	0.94	0.00045594	0.13	0.00114	0.80	-0.00454	0.29	0.00042362	0.89	0.00072089	0.79
9	-0.00316	0.33	-0.00257	0.49	-0.00493	0.09	0.00177	0.42	-0.00013967	0.57	-0.00421	0.16	-0.00391	0.31	-0.0043	0.10	0.00008207	0.97
10	-0.00112	0.73	0.0043	0.21	-0.00434	0.16	0.00322	0.06	0.00046503	0.13	-0.0003814	0.90	0.00532	0.12	-0.00334	0.23	0.00297	0.05

Table 3: Parameter estimate and p value for lagged sentiment measures produced by Bayes-Fisher and SVM classifiers

NN

lag	cc		oo		co		oc		spread		eccc		ecoo		ecco		ecoc	
1	0.00089198	0.85	-0.00005062	0.99	-0.00268	0.52	0.00357	0.12	0.0000888	0.65	-0.00134	0.78	0.00256	0.59	-0.0035	0.43	0.00215	0.28
2	-0.00018544	0.97	0.00534	0.16	-0.00272	0.56	0.00253	0.30	-0.00010813	0.53	0.0019	0.73	0.00566	0.08	-0.00043529	0.93	0.00238	0.26
3	0.00143	0.74	-0.00312	0.44	0.00271	0.45	-0.00128	0.60	0.00014492	0.67	0.0011	0.79	-0.00034131	0.92	0.00179	0.61	-0.00067498	0.75
4	0.00027233	0.95	0.00197	0.58	-0.00103	0.78	0.0013	0.52	-0.00032765	0.32	0.00419	0.28	0.00142	0.62	0.00282	0.44	0.00146	0.39
5	0.002	0.62	-0.00018654	0.96	0.00040267	0.91	0.00159	0.43	-0.00022739	0.30	0.00632	0.10	0.00273	0.35	0.00292	0.38	0.00333	0.05
6	-0.0005128	0.89	0.00377	0.27	-0.00178	0.55	0.00126	0.55	-0.00028003	0.44	0.00258	0.43	0.00298	0.37	0.0004823	0.85	0.00216	0.27
7	0.00208	0.61	-0.00218	0.56	0.00162	0.58	0.00045596	0.85	-0.00020352	0.65	0.00236	0.50	-0.00089887	0.81	0.00092307	0.71	0.00145	0.53
8	0.00198	0.61	0.00062015	0.86	-0.00163	0.57	0.00361	0.12	-0.00043306	0.19	0.00094201	0.80	0.00063296	0.86	-0.00207	0.43	0.00304	0.17
9	-0.00206	0.47	-0.00165	0.61	-0.00299	0.23	0.00093048	0.62	-0.00031224	0.24	-0.00176	0.49	-0.00182	0.58	-0.00254	0.25	0.00079024	0.65
10	-0.00188	0.51	-0.00003904	0.99	-0.00258	0.33	0.00069155	0.64	-0.00042125	0.21	-0.00263	0.31	-0.00152	0.61	-0.00235	0.33	-0.00024215	0.85

RNSE Data

lag	cc		oo		co		oc		spread		eccc		ecoo		ecco		ecoc	
1	0.00364	0.22	0.00077126	0.79	0.0031	0.24	0.00053878	0.73	0.00046926	0.04	0.00338	0.24	0.00076688	0.79	0.00302	0.25	0.00036811	0.79
2	0.00346	0.27	0.00081511	0.75	0.00001746	0.99	0.00345	0.03	0.00005651	0.71	0.00376	0.22	0.00047186	0.84	0.00125	0.64	0.00251	0.07
3	-0.00237	0.37	0.00050175	0.87	-0.00287	0.21	0.0005086	0.74	0.00037629	0.10	-0.00106	0.66	0.00098542	0.74	-0.00233	0.27	0.00129	0.35
4	-0.00194	0.52	0.00209	0.48	-0.00107	0.70	-0.0008709	0.55	0.00017053	0.44	-0.00167	0.58	0.00106	0.72	-0.00050783	0.86	-0.00117	0.39
5	0.00703	0.02	0.00419	0.11	0.00322	0.24	0.00382	0.01	-0.00002896	0.87	0.00491	0.11	0.00444	0.07	0.00134	0.63	0.00358	0.01
6	0.004	0.14	0.00366	0.20	-0.00012613	0.96	0.00413	0.01	-0.00010191	0.63	0.0029	0.23	0.00449	0.13	-0.00113	0.58	0.00402	<0.01
7	0.00155	0.61	0.00255	0.34	-0.00003749	0.99	0.00159	0.30	0.00022742	0.31	-0.00065236	0.83	0.00033338	0.89	-0.0019	0.48	0.00127	0.37
8	0.00126	0.65	0.00068839	0.81	0.00103	0.65	0.00023226	0.88	-0.00012542	0.54	-0.00267	0.30	-0.00051524	0.86	-0.00202	0.35	-0.00063563	0.65
9	0.0025	0.37	-0.00054481	0.83	0.00162	0.52	0.00087812	0.53	-0.00018352	0.38	0.00208	0.46	-0.0027	0.27	0.00147	0.57	0.00058709	0.65
10	0.00102	0.70	0.00355	0.16	-0.00107	0.64	0.00209	0.13	0.00006798	0.71	-0.00010018	0.97	0.00124	0.60	-0.00204	0.35	0.00194	0.11

Table 4: Parameter estimate and p value for lagged sentiment measures produced by NN classifier and from RNSE data

RNSE Data 2003-2008

lag	cc		oo		co		oc		spread		ecc		eoo		eco		eoc	
1	0.00192	*	0.00422	*	0.00069627	*	0.00122	*	-0.00022379	0.0013	0.00159	*	0.00315	*	0.00049224	*	0.00104	*
2	0.00079088	*	0.00094752	*	0.00053439	0.0001	0.0002565	0.0323	-0.00018928	0.0061	0.00032129	0.0376	0.00062796	0.0003	0.00018202	0.1377	0.00010481	0.3582
3	0.00054479	0.0023	0.00047096	0.0103	0.00059187	*	-0.00004708	0.6909	-0.00019863	0.0040	0.00013246	0.4290	0.00008823	0.6246	0.00022776	0.0621	-0.00011461	0.3805
4	0.00064604	0.0002	0.00063494	0.0005	0.00053741	*	0.00010863	0.3476	-0.00021883	0.0016	0.00015744	0.3067	0.00014738	0.3837	0.00014664	0.2254	-0.00002051	0.8586
5	0.00022031	0.2161	0.00029927	0.1003	0.00040797	0.0031	-0.00018766	0.1116	-0.00021297	0.0022	-0.00012185	0.4556	-0.0001969	0.2630	0.00012828	0.2879	-0.00028489	0.0234
6	0.00020479	0.2540	0.00048202	0.0086	0.00010578	0.4423	0.00009901	0.4102	-0.00022633	0.0012	-0.0001574	0.3664	0.00003479	0.8518	-0.00011089	0.3569	-0.00006389	0.6492
7	0.00025166	0.1574	0.00029764	0.1007	0.00007658	0.5771	0.00017509	0.1347	-0.00027038	0.0001	0.00019783	0.1838	-0.00008992	0.5809	0.00009277	0.4398	0.0000607	0.5710
8	0.00025473	0.1529	-0.00003164	0.8623	0.00039427	0.0042	-0.00013954	0.2397	-0.00025098	0.0003	0.00001777	0.9207	-0.0001344	0.4789	0.00022777	0.0596	-0.00023761	0.1003
9	0.000172	0.3371	0.00040752	0.0263	0.00011437	0.4064	0.00005763	0.6307	-0.0002348	0.0006	0.00003302	0.8616	0.00015286	0.4349	0.00003762	0.7555	-0.00002812	0.8524
10	0.00037959	0.0329	0.00017503	0.3333	0.00028868	0.0358	0.00009091	0.4365	-0.00024143	0.0006	0.00001263	0.9289	-0.00001409	0.9276	-0.00000507	0.9666	-0.00001812	0.8507
11	0.00032344	0.0748	0.00025458	0.1683	0.00037599	0.0060	-0.00005256	0.6695	-0.00021858	0.0018	-0.00003697	0.8519	-0.00038506	0.0578	0.0003613	0.0026	-0.00040459	0.0120
12	0.00022131	0.2169	0.00044602	0.0147	0.00018089	0.1862	0.00004042	0.7373	-0.00024489	0.0005	-0.00012861	0.4582	0.00015126	0.4126	0.00007943	0.5070	-0.00024499	0.0797
13	0.00041467	0.0187	0.00025657	0.1517	0.00036354	0.0076	0.00005112	0.6588	-0.00024455	0.0006	0.00010797	0.4786	0.00001232	0.9405	0.00016477	0.1667	-0.00007518	0.5089
14	0.00022718	0.2052	0.00050554	0.0055	0.00013618	0.3213	0.00009101	0.4504	-0.00025425	0.0003	0.00009126	0.5439	0.00014853	0.3607	0.00009409	0.4330	-0.00004179	0.7059
15	0.00003822	0.8286	-0.00000295	0.9869	0.00021168	0.1218	-0.00017346	0.1348	-0.00021112	0.0027	-0.00009376	0.5045	-0.00018682	0.2264	0.00017699	0.1395	-0.00029669	0.0023
16	0.0004305	0.0170	0.00029846	0.1048	0.00036374	0.0081	0.00006676	0.5825	-0.00021967	0.0017	0.00010826	0.5107	0.00013612	0.4411	0.00012293	0.3058	-0.00003922	0.7607
17	0.00028462	0.1203	0.00035585	0.0568	0.00025324	0.0647	0.00003138	0.8026	-0.00022816	0.0012	0.00021946	0.2327	0.00002064	0.9153	0.00029143	0.0153	-0.00009058	0.5510
18	0.00027666	0.1236	0.00028981	0.1121	0.00021893	0.1109	0.00005773	0.6304	-0.000231	0.0008	0.00009887	0.5301	0.00022895	0.1790	0.00013912	0.2442	-0.00005783	0.6331
19	0.00016156	0.3694	0.00015372	0.4024	0.00022874	0.0929	-0.00006717	0.5804	-0.00018249	0.0084	-0.00007894	0.6160	-0.00006326	0.7096	0.00008899	0.4579	-0.00017524	0.1478
20	0.00041741	0.0200	0.00028456	0.1162	0.00036731	0.0077	0.00005011	0.6739	-0.0001935	0.0054	0.00000844	0.9614	-0.00001209	0.9465	0.00009293	0.4407	-0.00009454	0.4753

Table 5: Parameter estimate and p value for lagged sentiment measures from RNSE data in period 2003 to 2008 (* p value ¡ 0.0001)

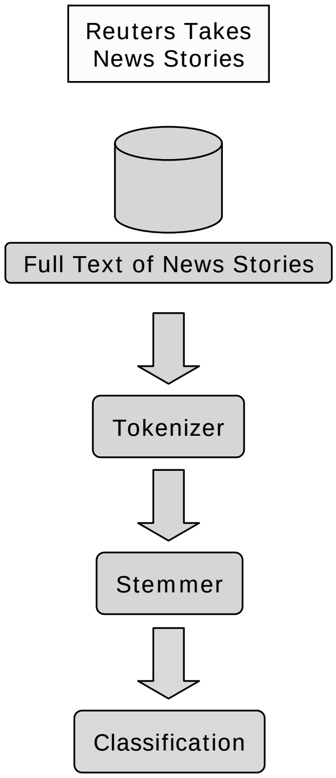

Figure 7: Transformations of news data

5 Conclusion

In an attempt to define a benchmark for performance of sentiment measures for financial texts, we propose the framework and present implementation of the system that enables easy deriving of financial indicators to be used in performance measure. As an illustration we offer a use-case analysis of four different approaches, although not giving any general conclusions about adequacy of particular approaches. It is shown that due to the inadequate data volume conclusions are vague, but the increase in the volume of analyzed data offers us a possibility to draw more robust conclusions.

Some of the future directions of the research would be extending the source base, including Compustat data on market capitalization and earnings of companies, and also extending calculated dataset, to offer wider variety of indicators which could be compared against candidate sentiment measure. More sophisticated statistical analysis should be performed on the final dataset to be able to give more decisive statements about general adequacy of text mining approaches in financial news. These could include using panel regression to control for inter-firm dependencies and applying significance test that would account for possible heteroscedasticity in the data.

References

Antweiler, W. and M. Z. Frank (2004): "Is All That Talk Just Noise? The Information Content of InternetStock Message Boards," *The Journal of Finance*, 59(3), pp. 1259–1294.

Bozic, C. (2009): "FINDS - Integrative services," in: *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on*, pp. 61–62.

Das, S. and M. Chen (2007): "Yahoo! for Amazon: Sentiment extraction from small talk on the web," *Management Science*, 53(9), pp. 1375–1388.

Fung, G. P. C., J. X. Yu, and W. Lam (2003): "Stock prediction: Integrating text mining approach using real-time news," in: *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pp. 395 – 402.

Gidófalvi, G. and C. Elkan (2003): "Using news articles to predict stock price movements," *Department of Computer Science and Engineering, University of California, San Diego*.

Hevner, A. R., S. T. March, J. Park, and S. Ram (2004): "Design Science in Information Systems Research," *MIS Quarterly*, 28(1), pp. 75–105.

Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan (2000): "Mining of Concurrent Text and Time-Series," in: *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mittermayer, M. and G. Knolmayer (2006a): "Text mining systems for market response to news: A survey," *Institute of Informations Systems, University of Bern*.

Mittermayer, M.-A. and G. F. Knolmayer (2006b): "NewsCATS: A News Categorization and Trading System," *Data Mining, IEEE International Conference on*, 0, pp. 1002–1007.

Pfrommer, J., C. Hubschneider, and S. Wenzel (2010): "Sentiment Analysis on Stock News using Historical Data and Machine Learning Algorithms," Term Paper, Karlsruhe Institute of Technology (KIT).

Schulz, A., M. Spiliopoulou, and K. Winkler (2003): "Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informationsüberlastung im Mobile Banking," *Wirtschaftsinformatik*, 2, pp. 181–200.

Tetlock, P. (2007): "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance*, 62(3).

Tetlock, P., M. Saar-Tsechansky, and S. Macskassy (2008): "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *Journal of Finance*, 63(3), pp. 1437–1467.

Thomas, J. (2003): *News and trading rules*, Ph.D. thesis, Carnegie Mellon University.

Wüthrich, B., D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam (1998): "Daily prediction of major stock indices from textual www data," *1998 IEEE International Conference on Systems, Man, and Cybernetics*.