News and stock markets: A survey on abnormal returns and prediction models

Mihail Minev, Christoph Schommer
ILIAS, University of Luxembourg

Theoharry Grammatikos
LSF, University of Luxembourg

Vast amount of news articles are published daily reflecting global topics. The stories represent information about events and expert opinions, which may trigger positive or negative expectations on the stock markets. The literature describes various methods for analyzing such correlations. In this paper we consider related approaches for tracking the impact of news on abnormal stock returns. In the first part we introduce studies with background in Finance. Primarily by applying statistical functions the works examine unusual price volatilities and explore possible sources and market conditions, e.g. biased investors, limited attention, macro-economic variables, country development state, et cetera. In the second part we present studies with background in Computer Science, which take advantage of historic news and the equivalent market values. By following the common learning paradigm the projects elaborate prototypes for trend and stock price prediction. In the current survey we evaluate leading approaches regarding the objectives, assumptions, input, techniques, and performance. Moreover we provide a comparison framework of the recent prototypes and identify gaps for future research.

Keywords: abnormal stock returns, news classification, prediction prototypes

1 Introduction

1.1 Topic Overview

Reducing the information overflow became a major challenge nowadays. The capacity to process unstructured data is limited not only by the time, but also by its complex structure. According to a leading press agency the amount of published news tripled since 2003. Thus the research community and commercial services are intensively looking for novel methods to quantify relevant topics and enable stock market predictions. In this context we face two main questions.

First, how to define which topics are indicatory. In many cases the relevance is determined by the recipients, which are the investors. This implies the requirement to analyze their reaction to a particular piece of information. Further the sources for abnormal market movements necessitate to be addressed. Pursuing these specific goals, the literature covers various research directions, i.e. investors psychology and limited attention, degree of market development, trading strategies, private and institutional investors, small and liquid stocks, information asymmetry and fundamental announcements. Relevant works are presented in the next chapter.

The second question is how to transform the quality data into a machine readable format. The idea is to quantify the information and gain insight applying techniques from the field of Computer Science, i.e. feature definition and extraction, term and phrases weighting, topic detection, and classification. Here, a matter of particular interest is the identification and measurement of correlations between news and stock returns. The determined relationships are used for the development of prediction prototypes. In this context we can distinct between applications, which claim to forecast (exact) stock prices, indices or trends and others, which aim to predict volatilities. Related works are evaluated in the third chapter.

Further purpose of this study is to identify potential research gaps. We analyze the models on the subject of the goals, target values, text sources, news types, number of documents and period considered, price temporal granularity, and applied machine learning algorithms. In order to determine the optimal features for a robust and well-performing prototype a comprehensive survey of all projects is presented in the forth chapter.

2 Tracking Stock Market Volatilities

In this section we evaluate studies with background in Finance, which track the sources for stock volatilities. Moreover the investors reactions to news and the subsequent exchange market movements are evaluated. Volatilities are specified and measured by the indicators of abnormal returns and high trading volume.

2.1 Study by Cutler et al.

"What moves stock prices" [CPS89] is a pioneer study from 1989 examining a broad range of news topics in relation to abnormal stock returns. *"The New York Times"* provided the articles for the time period 1941-1987. Cutler et al. claim that price fluctuations are measurably driven by factors, which are not only related to corporate control, earnings and regulatory policy. To identify outliers in time series the authors tracked price volatilities using vector autoregression. They considered monthly stock prices from 1926-1985 and annual returns from 1971-1986. Seven key macroeconomic figures from the United States

were included in the study.[1] Cutler et al. discovered that the related statistics are responsible for about 33% of the price variations. The most significant influence on stock returns was observed on information about the dividend payments, the industrial production, the inflation rate, and the volatility itself. Moreover the authors reviewed global non-economic stories, i.e. elections and military conflicts. However, these news types induced only weak impact on stock prices.

2.2 Study by Tetlock

In a work [Tet07] from 2007 Tetlock investigates the media influence on stock markets using the vector autoregression method (VAR). He analyzed the information summarized in the daily column *"Abreast of the Market"* from *"The Wall Street Journal"* over a 16 year period (1984-1999). Tetlock discovered that dominating negative news strongly influenced traders sentiment and outlook. Consequently the brokers triggered investments, which led to high volatility and low prices in a short term. The effect intensified, when the low returns generated even more negative media coverage. Moreover the study reveals the correlation between high pessimism and extreme trading volume. The results indicate to the significant role of investors psychology and bias for market volatilities. For the content analysis Tetlock integrated General Inquirer from Harvard. It is a mapping tool, which classifies terms based on a dictionary, i.e. *"Harvard IV-4"*.

2.3 Study by Mitra

In a collaborative[2] study Mitra et al. [Mit08] examine asset price volatilities in the context of portfolio risks. In order to anticipate future market developments and reduce uncertainty they consider a sentiment score for company news. The authors question the traditional multi-factor methods for risk estimation, i.e. macroeconomic, fundamental and statistical, due to their deficient feature to account quick condition changes. Consequently Mitra et al. propose a model for risk estimation, which further takes into account news sentiment and implied volatility. A distinction is drawn between unexpected news and time-fixed announcements with anticipated content. The quantitative input including a sentiment index is provided by RavenPack. Following the study results, portfolio risks are precisely estimated as a function of the strong relationship between news and market sentiment. However the study doesn't provide a casual assessment of volatilities by examining events and facts. Portfolio adjustments in correlation to "learned" experience from previous events would be an worthwhile extension to this work.

[1] The logarithm of real dividend payments; The logarithm of industrial production; The logarithm of real money supply; The nominal short-term interest rate; The nominal long-term interest rate; The monthly CPI inflation rate; The logarithm of the stock market volatility, defined as the average squared daily return on the S&P Composite Index within a month.

[2] Parties include RavenPack, OptiRisk Systems, Northfield, and Carisma.

2.4 Study by Fang and Peress

A study [FP09] from 2009 by Fang and Peress examines the relations between the news coverage of the four[3] major US newspapers and the cross-sectional dispersion of stock returns. The authors claim that stocks without related news outperform the opposite by 0.23% on average per month. Small stocks with low analyst and press attention score even higher with 0.65%-1% per month. The authors consider numerical figures like stock market characteristics, company size, book-to-market ratio and liquidity. Fand and Peress conclude that the information risk is the determinant factor for abnormal returns suppressing liquidity and investors behavior. This statement is also shared [EO04] by Easley and O'Hara in *"Information and the cost of capital"*.

2.5 Study by Chemmanur and Yan

Chemmanur and Yan [CY09] analyze the effects of advertising on stock returns by considering investors limited attention and their restricted cognitive capabilities. They discovered that increasing the advertising range and volume yields in a higher profit for the target year followed by an underperformance in the subsequent year. The effect is evident for stocks with larger costs of arbitrage, but also for small companies with low stock performance and popularity. The findings of Chemmanur and Yan comply with Barber and Odean [BO08], who claim that investors favor stocks, which have caught their attention, e.g. of companies presented in the media. The hypothesis rests on the observation about the humans limited attention, restricted to select only a few stocks among thousands. The authors state that personal preferences also influence the stock options. This is in contrast with the common market theory, where the determining factor is positive or negative information. The study was conducted on data from Dow Jones News Service.

2.6 Study by Talpsepp and Rieger

In a study [TR10] from 2010 including 49 countries Talpsepp and Rieger observe the relations between volatility asymmetry (a.k.a. "leverage effect"), news and the behavior of private investors. They examine the asset variations by implementing the asymmetric power model Garch. The results indicate that volatilities occur mostly in developed markets with high concentration of negative news. Countries with high market efficiency and economic advance lean to have a high level of volatility asymmetry, e.g. the United States ranking at first place followed by Greece, United Kingdom, Japan and Germany. Since the number of individual investors in these markets is significant, the trading activities triggered by investors' preferences and personal bias are also considerable. For instance, tracking search queries for terms like "recession", "inflation" and "oil prices" may indicate uncertainty. Further high volatility derived in the past after unanticipated events like the

[3]The Wall Street Journal, The New York Times, USA Today, and Washington Post.

financial crisis 2009 or the technology bubble crack 2001. The widespread hypothesis in the literature stating the substantial influence of the financial leverage could not be confirmed as the single trigger according to the authors. The survey of Talpsepp and Rieger correlates with the findings of Dzielinski et all [DRT09], who claim that private traders tend to overreact, intensifying the volatility.

2.7 Study by Dzielinski

In a recent paper [Dzi11] from 2011 Dzielinski compares the values of news and no-news stock returns. In this context he examines the correlation between sentiment-signed news, which may be positive, negative or neutral, and stock market volatilities. The study provides evidence for the efficient-market hypothesis (EMH) by discovering a fast market reaction to novel information, revealing priced sources of risk. Dzielinski claims that on days with dominating good news the returns are above-average and on those with more negative news the returns are below-average. Neutral articles are attested as non-relevant for the exchange markets. An interesting fact is the statistically significant price movement on the day before news release, which may be an evidence for information leaks. The results indicate, that market reactions are also motivated by factors like company size, book-to-market ratio and news coverage.

2.8 Study by Da et al.

In a recent article [DEG11] in *"The Journal of Finance"* Da et al. propose an innovative approach for assessment of investors attention by analyzing values from Google's Search Volume Index (SVI). The analysis is conducted with stocks from Russell 3000 for the years 2004-2008. The authors claim that capturing the extreme interest in a particular stock mostly leads to higher prices within the following two weeks. By contrast, initial public offerings (IPO) have usually a high first-day return.[4] The authors state that indirect proxies for investor attention are considerable, e.g. extreme returns, trading volume, news, advertising expenses, and price movements. But extreme returns and volume may be based on additional factors. Da et all claim Google's search engine statistics enable a direct measure of investors attention in relation to stock price volatilities.

3 Stock Market Prediction Models

The following prototypes attempt to imitate human reasoning. Commonly, investment decisions are triggered after an extensive study of all public information, e.g. news, opinions, interviews, press releases, et cetera. A key role for its systematic interpretation play the situation context, the relationships between the facts and the existing background knowledge of the recipient. However the quantification of latter figures is complex and error-prone.

[4] See Barber [BOZ09].

Consequently the implications of news on the stock markets are examined by discovering and modeling the price volatilities as functions of relevant announcements, i.e. fundamental and macro-economic. Analyzing historic course developments enable the identification of time periods with abnormal returns. The profitable opportunities are determined by the correlation with news and successively acquired as patterns. As a result, the prototypes facilitate information classification to one of the pre-defined categories, i.e. for prediction of exact price movements and trends as well as foreign exchange rates.

3.1 Study by Wüthrich et al.

In a pioneer study [WPL+98] from 1998 Wüthrich et al. attempted to forecast the closing values for five major equity indices – the Dow Jones Industrial Average (Dow), the Nikkei 225 (Nky), the Financial Times 100 Index (Ftse), the Hang Seng Index (His), and the Singapore Straits Index (Sti). The training set was composed of overnight articles from *"The Wall Street Journal"* and *"The Financial Times"*. The authors claim not only to consider the event major topic, but also its causal chain. However we are missing a strong evidence for a higher performance compared to the common approach in the study.

The application incorporated a model for counting and weighting terms (TFxCDF), described by Cho et al. [CWZ98]. The probabilistic rules were generated based on a previous work by Wütrich [Wüt97]. A handcrafted dictionary, provided by a financial expert, comprised about four hundred keyword records – pairs, triples, quadruples, or quintuples, e.g. "bond strong", "property weak", and "dow rebound". The final results claim an accuracy of at least 40% for all indices. This is slightly higher then the outcome by random guessing with 33%. However for the Hang Seng index the misclassification rose up to 28%. In a further experiment, rating only the Dow index, the prototype achieved a profit of 7,5%. Wüthrich et al. followed a daily trading strategy, closing all options by the end of the day. They took into account the daily closing price, which is different from the next day open price and may have led to inconsistencies. Trading cost were excluded from the evaluation.

3.2 Study by Lavrenko et al.

Lavrenko et al. [LSL+00] introduced in 2000 Ænalyst – a classification system developed to recommend news anticipated to influence the stock price direction. The outliers in time series are identified by using piecewise linear regression. Next, the abnormal returns are mapped against the time-stamped news stories and the relationships between them are analyzed. Lavrenko et al. used high frequency tick data in 10-minute intervals for 127 US stocks. In the training phase for each trend type (surge, slight+, plunges, slight−, others) the corresponding values for the patterns are calculated yielding in a language model. For instance, words like "loss", "shortfall", "bankruptcy" are expected to precede a negative reaction, while "merger", "acquisition", "alliance" a positive.

The data set contained 38.469 pre-categorized articles from Biz Yahoo! collected within a four months period. In the operational phase a news stream is monitored in real-time and labels are assigned to each text message. For the classification Naïve Bayes is implemented, the corpus is represented with the Bag-Of-Words method. By simulating a common trading scenario Lavrenko et al. achieve a modest gain of 0,23% per 10.000$ investment. Moreover the system achieved better results than the baseline vector-space (cosine similarity) approach. The authors assume that all documents issued five hours before the trend have equal importance, which is unrealistic. Transaction costs have not been taken into consideration. Parts of the study are based on previous research by Lavrenko et al. [LSL$^+$] and by Fawcett and Provost [FP99].

3.3 Study by Peramunetilleke and Wong

Predicting foreign exchange rates [PW02] is the main objective of the work by Peramunetilleke and Wong. They propose a model based on news headlines from the previous three hours, which by definition do not include background information and event details. The feature set comprised word pairs, records, and quadruples. Following thresholds were set for the three category classifier: up $> 0.023\%$, down $< -0.023\%$, and steady in-between. The handcrafted dictionary (by an expert) contained about 400 features in word groups, e.g. "US", "inflation", and "weak". However it is not publicly available and could not be compared to similar word lists. The training was conducted on a small input set dated from 21.09.1993-30.09.1993. The forecast rules were generated by experimenting with the three weighting techniques: boolean method, TFxCDF and TFxIDF. The results indicate an average accuracy of 48.6% using TFxCDF, which is much higher than results achieved by random guessing. For a very small training set a relatively high performance.

3.4 Study by Gidófalvi and Elkan

In a study [GE03] from 2003 Gidófalvi and Elkan experiment with extracting fundamental information and deriving quantitative indicators from financial news. From the beginning they discard ambiguous articles along with those outside of trading hours. To measure the effect on stock prices the authors describe a time interval, which captures the price movements 30 minutes before and after the story release. If a threshold value was overstepped during this period the news was classified as positive or negative. However the labeling process is in-transparent and insufficiently explained, e.g. missing the margin values. To evaluate the individual asset performance more precisely, the stocks are further divided in volatile and stable in relation to the Dow Jones Index. Gidófalvi and Elkan address also the case, where similar or identical information is republished. To avoid misclassification a similarity measure between the articles (24h) is proposed, eliminating such with values less then a threshold s. The feature set is build automatically over all terms and contains 1000 words with the highest mutual information. The records "sbc", "msft", "websphere",

"db", and "index" are ranked at the top. These terms are uncommon in comparison to features from other studies, which favor words with high degree of sentiment. The results indicate an average profit per trade of 0,01 % for the time interval $[-20, 0]$, but are not further interpreted and discussed.

3.5 Study by Mittermayer and Knolmayer

NewsCAT [MKK06] is an automated news categorization application aiming to forecast intraday price trends. The training set contains 989 company-specific press releases from different classes, e.g. sales reports, earnings, legal issues, et cetera. Mittermayer and Knolmayer used high frequency data (15-second-interval) to identify correlations with the S&P500. The prototype integrates three modules: a document processing engine, a categorization engine and a trading engine. The feature list is created automatically using initially Bag-of-Words. For the term weighting the authors incorporated the techniques Collection Term Frequency (CTF), Inverse Document Frequency (IDF) (default), and the combination CTFxIDF. Further functions supporting the automated text categorization in the study include Chi-squared (CHI), Information Gain (IG), and Odd's Ratio (OR). A handcrafted thesaurus[5], which is assumed to have a high relevance for stock price movements, is incorporated completing the automated feature detection. It contains not only single keywords, but also phrases and tuples of words/phrases. These features have higher priority in the final feature set, however the reason is not obvious. The input for the categorization engine are vectors with values acquired by Within-Document Frequency (WDF), IDF, WDFxIDF (default), or Boolification. The classifier (lSVM was used by default) claims to predict three categories with respect to the press releases – good, bad, and neutral. Other evaluated algorithms include Rocchio, kNN, and non-linear SVM. Mittermayer and Knolmayer reporter the highest performance for their prototype so far achieving 0.29 % per roundtrip. The study excludes press releases outside of the trading hours except for Nasdaq stocks. Transaction costs were also not considered. However, the results contradict partially with a similar study [SSW03] by Spiliopoulou et al. from 2003, where the authors could not identify measurable correlations primarily due to the high degree of noise in press releases.

3.6 Study by Schumaker and Chen

Stock price prediction is the goal of the *Arizona Financial Text System (AZFinText)* [SC09], developed by Schumaker and Chen at the Cleveland State University. The authors claim to predict exact stock prices within 20 minutes from news release (see also Gidofalvi in [Gid01]). The study tracked two main research questions: (1) *"What effect does GICS*[6] *partitioning of articles have on the prediction of stock price?"* and (2) *"How effective is*

[5] Not available online.

[6] Global Industry Classification Standard developed by Morgan Stanley.

a discrete prediction model versus the market and human traders?". The training set with financial news was provided by *Yahoo!*. The prototype assigned a weight to every single term in relation to historic prices from S&P500. Schumaker and Chen quantified the stories using a synthesis of linguistic and machine learning techniques. The authors considered four textual representations: Bag of Words, Noun Phrases, Named Entities and Proper Nouns. The later achieved the best scoring in combination with Support Vector Regression. Schumaker and Chen reduced the extracted features to a minimum of 3 occurrences per story, e.g. 8 terms for a 378 words document. This condition cuts down the noise for the training set, but concurrently increases the level of abstraction. *AZFinText* predicted a high directional accuracy of 71.18 % and a simulated trading return of 8.50 % outperforming selected human traders. One novelty in this approach was the story classification by industry, which has a positive effect on the forecast compared to individual news. Transaction costs were set to zero, and buy/sell were triggered only for the open/close price.

3.7 Study by Drury et al.

Drury et al. describe in [DTA11] a model, which combines automatically constructed dictionaries with manually implied rules, stories corresponding to abnormal stock returns, and a self-training algorithm. The goal of the work is to predict daily stock market index trends based on more than 300.000 general news. The rule classifier was designed to calculate either an event or a sentiment score (negative or positive) for the story headline based on a triple analysis, i.e. actor (who), verb/adjective, and object (profits, unemployment, etc). The rules were constructed as regular expressions in GATE. The authors further assume for each market index movement a particular sign in the time corresponding news stories (alignment strategy). The threshold of the second classifier for a positive or negative score was respectively set to FTSE-100>1.7 % and FTSE-100<2.11 %. Additionally, the outcome of both classifiers is compared and only the stories with identical labels are used for the training. Each time three models are induced from the training set and used for the text classification, until a stopping condition is reached (the paper does not provide details about this) or no new candidates are selected. The authors name it a self-training. The F-measure indicates the best performance for the later algorithm with a score of 0.84 for headlines. By contrast, the alignment strategy achieved only 0.57. This value may be a result of the noisy input including various news types. Interestingly the classification score based on the story text was only 0.71. The self-training algorithm achieved the highest return on headlines (47.2%) during the simulated trading. Though the explanatory power of this strategy is limited, due to the fact, that the system had access to all documents published on the particular single day. Unlike news are issued sequentially.

3.8 Study by Groth and Muntermann

[GM11] focus on risk management. The idea is to quantify news correlated with abnormal returns and to predict unusual price movements. The input consists of 423 regulatory-induced corporate disclosures, which are assumed to be a potential source of volatility signals. A high-frequency tick data is provided by *Thomson Reuters DataScope Tick History*. The documents are classified depending on the positive/negative relation of $ARISK$ and QUARTILE($ARISK_\tau$), where $ARISK_{[\tau_1,\tau_2]}$ is the abnormal risk with $[\tau_1, \tau_2] = [0, 15]$ and $[\tau_1, \tau_2] = [15, 30]$ minutes. The authors claim that this is the time interval, where the majority of the new information is incorporated into the stock prices and thus risk relevant. The resulting thresholds are respectively $ARISK_{\tau=15} = 0.281\%$ and $ARISK_{\tau=30} = 0.281\%$. Groth and Muntermann refer to [BKM02] for the text pre-processing task, which includes future extraction (algorithmic), feature selection (Chi-Squared-based), and feature representation (TFxIDF). They further use cost-sensitive learning methods, since the positive classification of risk intense news is more significant. A high precision of 100% (64.71%)[7] is achieved only in combination with a very low recall of 11.32% (31.13%) and a F_1 of 20.34 (42.04%). Unluckily the study does not provide details about the number of the considered features, their constitution and domain. Increasing the size of the feature set may achieve better classification, however it requires more training documents.

3.9 Study by Hagenau et al.

In a 2012 study [HLHN12] Hagenau et al. focus on the feature selection methodology aiming price prediction. The feature set is based on 2-word combinations (e.g. "result loss", "cell cancer", "reason delay") and includes nouns, articles, verbs, and other[8] terms. The class relevance is determined primarily by tracking the announcements (financial news) with the time corresponding stock values. In the first step the authors extract all words and represent them as N-Grams or word combinations. The training set is further extended by Noun Phrases selected with the Stanford Parser. Consequently the features, which are not relevant (Chi-Square based method) to a positive or negative market reaction, are excluded. The logarithm of the feature's frequency within a single text message is computed and adopted for the vector representation. The training is performed with 2/3 of the data using SVM, for which the authors claim[9] to produce the highest outcome. Without the sophisticated feature selection based on course relevance, the initial feature types performed similar (57.2 % – 58.6 %). The outcome rose remarkable by applying the Chi-Square based selection on the 2-words combinations. The project achieved with 65,1 % the best accuracy so far.

[7] Second best results using SVM.

[8] The authors did not provide more information for this category.

[9] The authors reference on tests with Neural Networks and Naïve Bayes.

4 Prediction Models Survey

In the previous section we described nine prototypes for stock markets prediction. We discovered numerous publications related to this topic, which indicates a high interest in the particular research field. Next, we conduct a cross-examine of the related works presented in the previous section. We further summarize the major prototype properties by elaborating a comparison framework. The results are introduced in Table 1 on page 14 and Table 2 on page 15.

4.1 Related Works

A survey [MK06] of studies, aiming stock markets prediction, was published in 2006 by Mittermayer and Knolmayer. They start with a brief overview of commonly applied text mining and classification techniques. In this context, the authors further present a procedure model for text categorization, which maps the major steps during the learning and the operational phase. Mittermayer and Knolmayer propose also a framework, where they describe the key characteristics of each project and discuss possible limitations. Next the properties of eight prototypes are extracted and compared including the forecasting objectives, the text mining parameters, the input data and the tests.

In an extra section Mittermayer and Knolmayer focus their attention on the stock markets prediction paradigm. They discuss the general adequacy of classical text mining approaches for the forecasting goal. One main concern for them are the market expectations and the requirement to capture these in the model. This would imply the development of a rule-based system and depreciate the text mining approach. The authors propose also, that future research should concentrate on incorporating the level of news anticipation. From our point of view this will further increase the quality of the training set by ignoring repetitive stories.

Mittermayer and Knolmayer identify the inclosed price-relevant information as an issue. Many financial news comprise such figures, which are problematic to interpret automatically. Numbers are mostly expressed as a ratio or a fraction of a whole and thus complex to render. On account of this challenge, a context-aware text analysis should decrease the level of abstraction. We think, that a certain amount of incorporated background knowledge will enable a systematic entity comparison, e.g. company data from the previous quarter.

4.2 Prediction Prototype Framework

Table 1 and Table 2 are constructed with eight attributes like specified. We grouped the studies build upon their goal, i.e. prediction of price trends (Table 1) and the rest (Table 2). For more clarity we use two distinct representations, but keep the attributes identical. The first and the second columns are dedicated to the project goal and the target values for which one aims a prediction, e.g. DAX, FTS-100, or individual stocks.

Next are the values for the sources of the texts (Yahoo! Finance, Thomson Reuters, etc.) and the source types, e.g. financial news, press releases, or general articles. In the fifth column we list the considered time period and the number of documents in the training set (if available). The sixth part specifies the price frequency used for the story correlation with the financial markets. The attribute "Techniques" lists all the algorithms, methods, and classifiers evaluated in the particular study. If we revealed information about the highest/lowest performer with indicate it in brackets. The last column comprises all other important prototype characteristics, e.g. automatic or handcrafted dictionary, number of features, classification categories, and evaluation metrics.

Although all studies aim prediction of figures related to the stock markets, their goals are little diffuse. The majority, and in particular four (see Table 1) of the nine studies, aim to forecast stock price trends. Two of them propose a 3-category, one a 2-category and one a 5-category classifier. However, the classes may be summarized twofold – positive or negative. The neutral is not of interest since the data is not market relevant and thus applicable for prediction. The four studies adopt an algorithmic feature extraction. Though Mittermayer and Knolmayer adopt additionally a handcrafted dictionary containing 423 terms. Naïve Bayes is the preferred classifier in two of the projects, whereas the other two reached the highest scores with nlSVM and SVM respectively. The highest accuracy of 65,1 % was achieved by Hagenau et al., followed by Mittermayer and Knolmayer with 45 %, and Gidófalvi and Elkan with 40 %. Lavrenko et al. did not provide comparable measurements.

Table 2 summarizes the properties of the remaining five studies. They pursue slightly divergent goals, but all of them focus on predicting stock markets figures. Schumaker/Chen and Wütrich et al. are the two studies, which claim to predict exact stock prices and indices respectively. The later use general news, e.g. economic, political and interviews and a price granularity of one day. These type of non-financial news are considered to trigger volatilities and are also analyzed by Peramunetilleke/Wong and Drury et al. Whereas Schumaker/Chen process company related financial news classified by industry. In comparison they achieve a higher accuracy with 58.2 % compared to the 44 % by Wütrich et al. Incorporating the largest data set of all projects Drury et al. aim to predict the FTS-100 index. By contrast they use a self-training algorithm and achieve a 47.2 % accuracy using only the headlines. This is not the highest score indicating that big data does not necessary imply better prediction. However, we acknowledge the results are barely equipollent, since each projects incorporates unique data and the source code is omitted.

Groth and Muntermann rely on automatic feature extraction as also six of the other studies do. The training data set contains less than 423 corporate disclosures. The prediction goal is abnormal price movements, which are indeed essential for the risk management domain. For the term weighting they implement a Chi-Squared based method as also Hagenau et al. do. One advantage of it is the accurate term weighting, due to the correlation step with the stock prices. An alternative and similar technique is information gain. The accuracy of the model is 78.49 % in relation to a very low recall of 31.13 %.

One of the two studies incorporating a single handcrafted dictionary aims to predict exchange rate trends. The feature set proposed by Peramunetilleke/Wong contains 400 terms of two to five words and has a similar volume to the one used by Wütrich et al. with 423 tuples. The dictionaries are elaborated by a financial expert, which may be a limitation. Even though the prototype yields an accuracy of 50 % and thus ranges in the middle of the field.

The majority of the prototypes are build for processing documents in English and focus on the prediction of the United States stock market. This may be due to the large amount of public available data (see Linguistic Data Consortium), but may be also induced by the language complexity. Moreover many of the text processing and analysis tools function properly only with English sources. On the other hand facilitating baseline surveys on the incorporated feature sets is beneficial.

For the story classification the most applied methods are Naïve Bayes and variations of SVM, e.g. nlSVM, lSVM, SVR. While neural networks and the k-nearest neighbor algorithm (k-NN) are used by just a few authors. Studies, which evaluate a number of machine learning approaches, claim higher-ranking output for SVM. However, this may be a result of the parametrization or the underlying training set and requires more evidences.

Various weighting techniques exist, which determine the importance of terms to a document in the news collection. Many authors favor the term frequency–inverse document frequency (TFxIDF) and the term frequency-cumulative distribution function (TFxCDF). Proper alternatives for these are the Chi-Square based methods (Hagenau et al.; Groth/-Muntermann) and information gain (Groth/Muntermann; Mittermayer/Knolmayer). By contrast to TFxIDF and TFxCDF, these weighting methods consider also the exact stock courses.

Almost none of the studies assay trading costs, besides Hagenau et al. Even though we expect them to influence the results of the simulated trading. Transaction costs include brokers' commissions and spreads but also the limited market liquidity. Due to the taxes incompatibility, e.g. by industry, country, and stock exchange, a more practical value is required. Hagenau et al. calculate with transaction costs of 0.1 %. On the other hand, omitting the extra fees renders more conclusive results.

Forecast	Target Value(s)	Text Source(s)	News Type	Period & Training Data	Price Temporal Granularity	Techniques	Other	Study
Stock Price Trends	HDAX	DGAP and EuroAdhoc	German/UK Adhoc messages	1998 – 2012; 7240 (DGAP)	daily open/close	Porter Stemmer, N-Grams, Noun Phrases (Stanford Parser), Chi-Square, SVM (best)	Automatic feature extraction; 2-category model; 2-word combinations; 65,1 % accuracy	[HLHN12]
Stock Price Trends	S&P 500	PRNewswire	Press Releases (company-specific)	01.04.2002 – 31.12.2002; 989	15 sec	Bag-of-Words; CTF(best); IDF; CTFxIDF; WDF; CHI; IG; OR; WDFxIDF; k-NN; nlSVM(best); lSVM	automatic feature extraction and handcrafted thesaurus; tuple (terms); 3-category model; 45 % accuracy	[MKK06]
Stock Price Trends	Dow Jones	unknown	financial news articles	Jul 26, 2001 – Mar 16, 2002; 3200-3400	1 min	Rainbow Naïve Bayes classifier package; Witten-bell smoothing method; Regression	3-category model; automatic feature extraction using MI (1000 single words with highest mutual information); 40 % accuracy	[GE03]
Stock Price Trends	127 U.S. stocks	Yahoo! Finance	online news articles; company related stories	Oct 1999 – Feb 2000, 38469	10 min	TFxIDF; Naïve Bayes	5-category model; automatic feature extraction; terms; analysis 5-10h before trend	[LSL$^+$00]

Table 1: Related works in Computer Science – Prediction of Stock Price Trends

Forecast	Target Value(s)	Text Source(s)	News Type	Period & Training Data	Price Temporal Granularity	Techniques	Other	Study
Volatility	DAX	DGAP	regulatory-induced corporate disclosures	Aug 1, 2003 – Jul 31, 2005; 423 (in total)	15 min	feature selection based on Chi-Squared (best) and Information Gain; feature representation – TFxIDF; Naïve Bayes (worse); kNN; Neural Network; SVM (libsvm, best)	automatic feature extraction; 2-category model; Precision, Recall, Accuracy, F_1-measure, Simulated trading; 78.49 % accuracy	[GM11]
Stock Market Index Trends	FTS-100	unknown source (RSS feeds)	various news types	Oct, 2008 – Jul, 2012; >300.000 (in total)	daily open/close	Regular expressions; self-training algorithm; voting model	2-category classifier; automatic feature extraction; manual rule classifier; alignment strategy; event/sentiment score; F-Measure; Simulated trading; 47.2 % accuracy	[DTA11]
Stock Prices (exact)	S&P500	Yahoo! Finance	online financial news articles; company related; classification by industry – GICS	Oct 26, 2005 – Nov 28, 2005; 2809	1 min	Bag-of-Words; Noun Phrases; Named Entities; Proper Nouns (best); SVR (in WEKA)	automatic feature extraction; eval. on measures of Closeness, Directional Accuracy and Simulated Trading; GICS Analysis; 58,2 % accuracy	[SC09]
Exchange Rate Trends	USD/DEM USD/JPY	Reuters	financial, political, general economic news headlines	Sep 21, 1993 – Sep 30, 1993; est. 2400 (60h with 40 headlines each)	1 hour	Decision Rules; Boolean method; TFxIDF, TFxCDF (best)	3-category model; handcrafted dictionary from a trading expert with 400 features (two to five words); 50 % accuracy	[PW02]
Stock Market Index (exact)	DJ; Nikkei; FTSE; HS; ST	WSJ, FT	online economic and political news, analysis results and interviews	Dec 1997 – Mar 1998 (100 trading days)	daily close	TFxCDF; Neural Network with backpropagation; FFNN; k-NN (best)	3-category model; handcrafted dictionary with 423 features; tuple of words ; 44 % accuracy	[WPL+98]

Table 2: Related works in Computer Science – Prediction of Stock Market Figures

5 Discussion

In general market prediction models are limited per se, due to the nature of the stock exchange itself. Course movements are not directly induced by information, whereas by human or algorithmic traders. Each investor incorporates a coherent strategy, which is a result of manifold factors. In common is the pursuit to outperform the other agents and be the first to identify the profitable opportunity. This goal implies the conclusion, that if an optimal strategy exists, it will be determined and reflected by every rational market participant. However this approach contradicts with the way stock markets function. If the majority follows the predetermined superlative strategy, it will become less profitable and thus obsolete in a very short time.

The suboptimal results from the simulated trading may be partially reflected by the latter aspects. Even though document classification with respect to stock markets remains challenging. Essential is the interpretation of news and information respectively. Future applications may not only focus on single stories, but also consider the context and the topic relationships. In great demand is a more accurate identification of relevant news, i.e. with a high market significance.

The data volume is also a capital aspect. A big training set enables more flexible models and thus supports unusual experiments. For instance, performance tests with the feature selection, extraction and representation. Moreover heterogeneous validation sets may positively influence the classifier accuracy.

References

[BKM02] Heide Brücher, Gerhard Knolmayer, and Marc-Andre Mittermayer. Document classification methods for organizing explicit knowledge. In *Proceedings of the Third European Conference on Organizational Knowledge, Learning, and Capabilities*, Athens, Greece, 2002. University of Bern.

[BO08] Brad M. Barber and Terrance Odean. All that Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies, Vol. 21, No. 2, pp. 785-818, 2008*, 2008.

[BOZ09] Brad M. Barber, Terrance Odean, and Ning Zhu. Do retail trades move markets? *Review of Financial Studies*, 22(1):151–186, 2009.

[CPS89] David M. Cutler, James M. Poterba, and Lawrence H. Summers. What Moves Stock Prices? *SSRN eLibrary*, 1989.

[CWZ98] V. Cho, B. Wüthrich, and J. Zhang. Text processing for classification, 1998.

[CY09] Thomas J. Chemmanur and An Yan. Advertising, attention, and stock returns. *SSRN eLibrary*, 2009.

[DEG11] Zhi Da, Joseph Engelberg, and Pengjie Gao. In search of attention. *The Journal of Finance*, 66(5):1461–1499, 2011.

[DRT09] Michal Dzielinski, Marc Oliver Rieger, and Tonn Talpsepp. *Volatility asymmetry, news and private investors*. Wiley, 2009.

[DTA11] B. Drury, L. Torgo, and J.J. Almeida. Classifying news stories to estimate the direction of a stock market index. In *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*, pages 1 –4, june 2011.

[Dzi11] Michal Dzielinski. News sensitivity and the cross-section of stock returns. NCCR FINRISK, Number: 719, `http://www.nccr-finrisk.uzh.ch/wps.php?action=query&id=719`, 2011.

[EO04] David Easley and Maureen O'hara. Information and the cost of capital. *Journal of Finance*, 59(4):1553–1583, 08 2004.

[FP99] Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. In *In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 53–62, 1999.

[FP09] Lily Fang and Joel Peress. Media coverage and the cross-section of stock returns. *Journal of Finance*, 64(5):2023–2052, October 2009.

[GE03] Győző Gidófalvi and Charles Elkan. Using news articles to predict stock price movements. Technical report, Department of Computer Science and Engineering, University of California, 2003.

[Gid01] Győző Gidófalvi. Using news articles to predict stock price movements, 2001.

[GM11] Sven S. Groth and Jan Muntermann. An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50(4):680 – 691, 2011. ¡ce:title¿Enterprise Risk and Security Management: Data, Text and Web Mining¡/ce:title¿.

[HLHN12] Michael Hagenau, Michael Liebmann, Markus Hedwig, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-specific features. *Hawaii International Conference on System Sciences*, 0:1040–1049, 2012.

[LSL+] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *In proceedings of the 6 th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 37–44.

[LSL+00] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, James Allan, and James Allan. Language models for financial news recommendation. pages 389–396, 2000.

[Mit08] Leela Mitra. Equity portfolio risk (volatility) estimation using market information and sentiment equity portfolio risk (volatility) estimation using market information and sentiment. *Analysis*, pages 2–13, 2008.

[MK06] Marc-André Mittermayer and Gerhard F. Knolmayer. Text mining systems for market response to news: A survey. Working Paper No 184, August 2006.

[MKK06] Marc-André Mittermayer, Gerhard F. Knolmayer, and Gerhard F. Knolmayer. Newscats: A news categorization and trading system. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 1002–1007, 2006.

[PW02] Desh Peramunetilleke and Raymond K. Wong. Currency exchange rate forecasting from news headlines. *Aust. Comput. Sci. Commun.*, 24(2):131–139, January 2002.

[SC09] Robert P. Schumaker and Hsinchun Chen. A quantitative stock prediction system based on financial news. pages 571–583, 2009.

[SSW03] Myra Spiliopoulou, A Schulz, and K Winkler. *Text Mining an der Börse: Einfluss von Ad-hoc-Mitteilungen auf die Kursentwicklung*, pages 215–228. 2003.

[Tet07] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168, 2007.

[TR10] Tõnn Talpsepp and Marc Oliver Rieger. Explaining asymmetric volatility around the world. *Journal of Empirical Finance*, 17(5):938 – 956, 2010.

[WPL+98] Beat Wüthrich, D. Permunetilleke, S. Leung, Vincent Cho, Jian Zhang, and W. Lam. Daily prediction of major stock indices from textual www data. In *KDD*, pages 364–368, 1998.

[wSGSS04] Young woo Seo, Joseph Giampapa, Katia Sycara, and Katia Sycara. Financial news analysis for intelligent portfolio management. 2004.

[Wüt97] Beat Wüthrich. Discovering probabilistic decision rules. *Int. Syst. in Accounting, Finance and Management*, 6(4):269–277, 1997.

[YJDS06] Ting Yu, Tony Jan, John K. Debenham, and Simeon J. Simoff. Classify unexpected news impacts to stock price by incorporating time series analysis into support vector machine. In *IJCNN'06*, pages 2993–2998, 2006.