Decision Support Systems 62 (2014) 32–42

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

How to enable automated trading engines to cope with news-related liquidity shocks? Extracting signals from unstructured data

Sven S. Groth, Michael Siering *, Peter Gomber

Goethe University Frankfurt, Grüneburgplatz 1, 60323 Frankfurt, Germany

ARTICLE INFO

Article history:
Received 9 July 2012
Received in revised form 10 February 2014
Accepted 6 March 2014
Available online 16 March 2014

Keywords:
Automated trading
Liquidity
Forecasting
Text mining
e-Finance
Simulation

ABSTRACT

Financial markets are characterised by high levels of complexity and non-linearity. Information systems have often been applied to support investors by forecasting price changes in securities markets. In addition to the asset price, liquidity represents another financial variable that has a high relevance for investors because it constitutes a main determinant of total transaction costs. Previous research has shown that the level of liquidity is affected by the publication of corporate disclosures. To derive an optimal order execution strategy that minimises the transaction costs, investors as well as automated trading engines must be able to anticipate changes in the available market liquidity. However, there is no research on how to forecast the impact of corporate disclosures on market liquidity. Therefore, we propose an IT artefact that allows automated trading engines to appropriately react to news-related liquidity shocks. The system indicates whether the publication of a regulatory corporate disclosure will be followed by a positive liquidity shock, i.e., lower transaction costs compared to historical levels. Utilising text mining techniques, the content of the corporate disclosures is analysed to generate a trading signal. Furthermore, the trading signal is evaluated within a simulation-based use case that considers English and German corporate disclosures and is shown to be of economic value.

1. Introduction

Decision-making in the financial domain represents a challenging task because financial markets are characterised by high levels of complexity and non-linearity [8]. These market characteristics make it difficult for decision-makers to quickly adjust their strategies in cases of company-related or economic events. Here, information systems play a crucial role in supporting human and computer-based decision-makers alike.

Today, the group of computer-based *automated*[1] *traders* generates already approximately one-half of the trading activity on major European markets such as Deutsche Börse's Xetra, and the percentage share continues to grow [7]. Computer-based *automated traders* "emulate a broker's core competence of slicing a large order into a multiplicity of smaller orders and of timing these to minimise the market impact" [13]. The decision on the investment or portfolio allocation itself is performed by the respective portfolio manager at a fund management company, and the primary task of *automated traders* is to execute the orders that are received from these fund management companies or institutional investors at the best available conditions.

To determine an optimal execution strategy for a pre-defined execution time period, i.e., to achieve the best available conditions, *automated traders* must handle the trade-off between the evolving transaction cost components in the order's execution. Costs that are generated while implementing investment decisions can generally be divided into two broad categories: First, there are explicit costs, such as commissions, fees, and taxes. Second, there are implicit costs, such as market impact, timing costs, and opportunity costs [1]. Especially for large trades, implicit transaction costs are mostly much larger than explicit transaction costs. Liquidity constitutes the main determinant of implicit transaction costs: if the number of shares that other market participants are willing to trade at a given limit is reduced, then the market impact of an order is increased [9]. Thus, liquidity and implicit transaction costs represent two sides of the same coin [41]: The higher the liquidity is, the lower the implicit transaction costs, and vice versa. Thus, "asset managers and ordinary investors care about liquidity insofar as it affects the return on their investments, simply because illiquid securities cost more to buy, and sell for less" [12]. Therefore, to derive an optimal execution strategy and to minimise the associated transaction costs, the ability to forecast future liquidity levels is very important.

Bikker et al. [1], however, conclude that "forecasting market impact costs appears notoriously difficult and traditional methods fail". Moreover, Domowitz and Yegerman [10] find that the execution quality of *automated traders* is inferior to the executions that are handled by brokers. One possible reason for this observation might be the fact that the currently employed models are solely based on purely

* Corresponding author. Tel.: +49 69 798 33832; fax: +49 69 798 35007.
E-mail addresses: sgroth@wiwi.uni-frankfurt.de (S.S. Groth), siering@wiwi.uni-frankfurt.de (M. Siering), gomber@wiwi.uni-frankfurt.de (P. Gomber).

[1] The terms *automated trading* and *algorithmic trading* are used interchangeably in this study.

http://dx.doi.org/10.1016/j.dss.2014.03.002
0167-9236/

quantitative data input. Existing (academic) models and strategies largely neglect one of the most important sources of information, which is unstructured qualitative data (i.e., news) [6]. If, for example, a listed company issues an unanticipated regulatory ad hoc disclosure, then *automated traders* cannot react sufficiently fast simply because they cannot analyse the content. Because unanticipated news, by definition, is very unlikely to be reflected in quantitative time series data prior to its publication date, *automated traders* can respond only to other (human) market participants' reactions.

Against this background, our research goal is to investigate whether and how unstructured qualitative data can be used as input for *automated trading engines*. We are especially interested in whether useful information can be extracted automatically from qualitative data to predict future levels of liquidity after the publication of corporate disclosures. For extracting such information, text mining techniques are utilised. Given that decision support in the financial domain is very challenging [8], we especially call the reader's attention to those domain-specific issues that require an adjustment of standard knowledge discovery approaches, to investigate and emphasise the economic relevance of the proposed system. We particularise the knowledge discovery in databases (KDD) process proposed by Fayyad et al. [11] by domain-specific customisations, such as the application of an event study for *data understanding* and a novel evaluation scenario in the form of a trading simulation. Moreover, we investigate the role of language within the proposed text mining framework. In other words, we enquire whether the proposed text mining system is sensitive to the language of the input text.

Thus, we contribute to the literature on financial text mining by proposing an IT artefact to forecast the liquidity impact of corporate disclosures. In contrast to previous studies that mainly focus on forecasting the stock price impact of financial news, we concentrate on the most important criterion of market quality, which influences the highest cost component in trading, i.e., the implicit transaction costs [25]. Furthermore, we enhance previous research by focusing on the economic relevance of financial text mining systems by extending the KDD process by Fayyad et al. [11] by means of an event study and a novel simulation. In this way, the simulation aims at evaluating the economic value of the proposed system and extends previous studies by accounting for the timing of the orders. Finally, we investigate whether the results differ for different language inputs.

The remainder of this paper is structured as follows: First, we present related financial and information systems research. Second, we describe the study setup, which is based on the KDD process proposed by Fayyad et al. [11]. Third, the adjusted KDD process is applied to the above-described *automated traders'* use case. In this way, the dataset that is used in this study is described, and the liquidity impact of the publication of regulatory corporate disclosures is investigated, to enhance the data understanding. Fourth, building on these insights, we propose a text mining approach that predicts the liquidity impact of the ad hoc news. The classification quality is evaluated with respect to both the *classic* model evaluation metrics and domain-specific *simulation-based* model evaluation. Finally, we present our conclusions.

2. Related work

2.1. Financial text mining

There are several studies that apply text mining techniques in financial markets to find patterns in text that can serve for predictions. In this context, Mittermayer and Knolmayer [35] provide a good survey on existing text mining systems: Since the time that Wuthrich et al. [46] proposed one of the first financial text mining applications, systems have been refined by focusing on aspects such as intraday data, (new) data mining techniques, other forecasting objects (stock prices, exchange rates, volatility), news types (ad hoc news), and novel evaluation methods [17]. The prediction of company-specific liquidity levels in general and the prediction of the liquidity impact of regulatory corporate disclosures in particular have – to our knowledge – not yet been addressed by utilising text mining techniques. Additionally, the language of input texts has been addressed in the literature [3] but not with regard to the financial industry. Loughran and McDonald [31], however, have highlighted that financial texts require domain-specific knowledge and interpretation.

As a consequence, our approach builds upon and extends this literature by using intraday high-frequency data on the new forecasting object *liquidity*. We additionally concentrate on a certain news type and develop a novel domain-specific evaluation metric.

2.2. Liquidity impact of corporate disclosures

Liquidity refers to the possibility of buying or selling an asset immediately without adversely affecting the price, and it is seen to be the most important aspect of market quality [14]. Liquidity is composed of two key dimensions that are relevant in our context: A market has *breadth* when the best buy and sell orders exist in substantial volume (see Fig. 1). Additionally, a market has *depth* when there are orders in substantial volume in the closest neighbourhood to the best bid and best ask limits [41]. To evaluate breadth and depth, an open limit order book that displays the cumulated buy and sell orders can be used. Two exemplary open limit order books are shown in Fig. 1. On the left of each order book, the buy orders (bid) are displayed, whereas on the right, the sell orders (ask) are indicated; both have order limits and available quantity at their respective limits.

In general, a lower level of market breadth or market depth (and, consequently, a lower level of liquidity) is disadvantageous for market participants because these levels would implicate higher implicit transaction costs. Referring to the order book examples that are displayed above, an investor who is willing to buy a quantity of 145 stocks would have to pay 53.00 for each stock in order book situation 1. However, if there were only 100 shares available at the best ask of 53.00, the investor would have to buy additional 45 shares for 54.00 to obtain the desired quantity (order book situation 2). The resulting average price of 53.31 per share would be worse than it has been in the more liquid market. As a result, in order book situation 2, market participants would have to bear an increased amount of implicit transaction costs.

To summarise, liquidity refers to the bids and offers that are provided in the market and are listed in the order book, i.e., it is shown ex-ante which trading opportunities are available for traders. In contrast, prices and trading volume result from the liquidity demand, i.e., they reflect past trades, and because they are ex-post, they are not relevant for decisions that are related to the timing of orders. Therefore, in contrast to existing research, we focus on the relevant pre-trade decision criterion (i.e., the liquidity) instead of focusing on the impact of corporate disclosures on prices ex-post.

In previous studies, the effect of information arrival on company-specific liquidity levels has already been analysed (e.g., [28]). Most of the contributions, however, merely concentrate on one event type, such as earnings announcements [26] or dividend announcements [15]. Moreover, authors use liquidity measures that do not allow a complete analysis of transaction costs because they account only for the market breadth [39]. Finally, other studies are based on very few events and/or short time periods [4,14]. We address all of the above shortcomings in this paper to extend the previous literature: We analyse a comparatively large dataset that is composed of several regulatory news types. Being provided with high-frequency order book data, we apply a liquidity measure that is especially suitable for the estimation of liquidity (i.e., the implicit transaction costs). Finally, we focus on short-term intraday liquidity effects.

a) *Order Book Situation 1 (Liquid Market)*

Market breadth

Market depth

BID		ASK	
Quantity	Limit	Limit	Quantity
150	52.00	53.00	145
200	51.00	54.00	65
50	50.00	55.00	100
75	49.00	56.00	110
100	48.00	57.00	200
…	…	…	…

b) *Order Book Situation 2 (Less Liquid Market)*

Market breadth

Market depth

BID		ASK	
Quantity	Limit	Limit	Quantity
60	52.00	53.00	100
0	51.00	54.00	55
70	50.00	55.00	0
0	49.00	56.00	5
15	48.00	57.00	35
…	…	…	…

Fig. 1. Two exemplary open limit order books.

3. Research approach

Generally, data mining aims at discovering useful patterns in data that can serve for predictions [11]. However, for a successful application of machine learning techniques, additional steps such as domain and data understanding as well as data reduction and pre-processing are crucial, too [18]. To account for these steps properly, our study is grounded on the KDD process model that is proposed by Fayyad et al. [11]. In comparison to other process models, it is considered to be the most suitable for data mining projects that require substantial data pre-processing activities [27]. Additionally, it has gained a large amount of attention within the literature on related data mining studies [27].

Fig. 2 shows our research approach based on the KDD process model by Fayyad et al. [11]. In our study, the phases *understanding the application domain* and *creating a target dataset* are conducted in parallel: On the one hand, we acquire a dataset that is composed of ad hoc disclosures, stock prices and order book data; on the other hand, the literature review presented in Section 2 and the event study presented in Section 5 help us to understand the liquidity impact of ad hoc disclosures. Thereafter, the phases of *data cleaning & pre-processing* as well as *data reduction* are performed. Consequently, a *data mining method* and an appropriate *data mining algorithm* are selected to conduct data mining. Within our study, we choose Support Vector Machine (*SVM*) as an appropriate data mining algorithm for classification. Finally, it is essential that the results of the data mining phase are analysed [2]. Therefore, the *results are interpreted* by means of *classic* model evaluation, and the *discovered knowledge* is applied within a *simulation-based* evaluation.

4. Creating a target dataset

The news dataset at hand is composed of ad hoc disclosures that are published by *Deutsche Gesellschaft für Ad-hoc-Publizität* (DGAP) on behalf of the companies admitted to trading on an organised market in Germany. To fulfil the legal requirements, these companies must publish immediately any insider information or other information that is highly relevant to investors. We concentrate on this news type because the disclosures are expected to primarily contain new information and event studies have shown that these are often followed by abnormal stock returns [37].

We include in our dataset only those corporate disclosures that were published during exchange trading hours because we focus on intraday liquidity effects. Our event study analysis requires market data 15 min after/prior to an ad hoc publication as input; the earliest time for inclusion into the dataset is 9:15 a.m., and the latest time is 5:15 p.m. Moreover, we concentrate on companies that were members of one of the following German stock indices at the disclosure publication date: DAX (large-capitalisation stocks), MDAX (medium-capitalisation stocks), and SDAX (small-capitalisation stocks). A total of 71 disclosures were excluded from the dataset because of time window conflicts; disclosures were not included in the dataset if there was another disclosure from the same company during the N = 30 days prior to publication. This setup ensures that potential liquidity effects can be analysed in isolation. Consequently, the final dataset comprises 415 ad hoc disclosures that were published between 2006-01-31 and 2009-07-22. Furthermore, the observed liquidity effects remain robust if a time period of 15 days is used to exclude the confounding events.

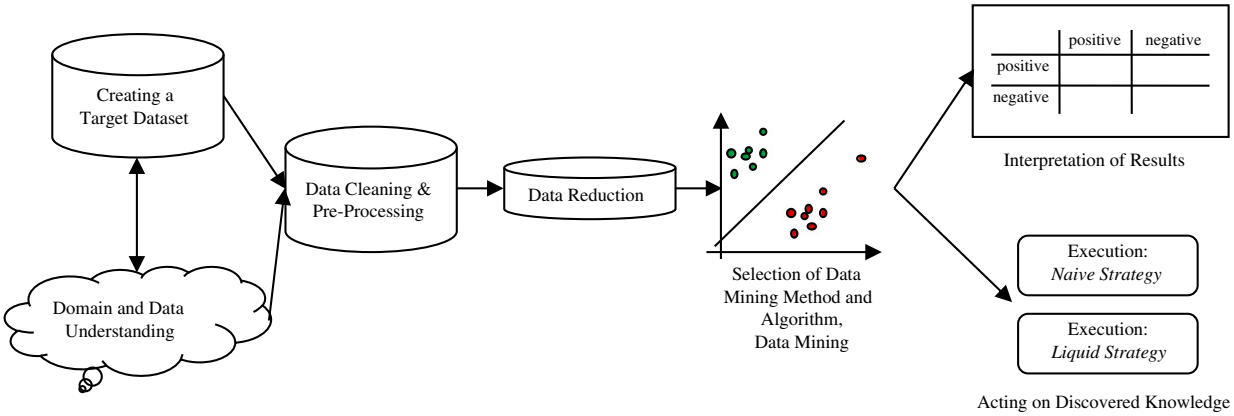

Fig. 2. Research approach based on the KDD process by Fayyad et al. [11].

Table 1
Median $ACRT(v)_{[t1,t2]}$ sorted by volumes v.

	Time period					
Metric	[−30, −15]	[−15, 0]	[0, 15]	[15, 30]	[30, 45]	[45, 60]
$ACRT(v = median)_{[t1,t2]}$	−0.045	−0.018	0.307***	0.140***	0.102***	0.048***
$ACRT(v = 4^*median)_{[t1,t2]}$	−0.030	−0.023	0.273***	0.115***	0.066***	0.037***
$ACRT(v = 25\ k)_{[t1,t2]}$	−0.021	−0.021	0.280***	0.116***	0.060***	0.024***

***/**/* indicate significance at the 1%/5%/10%-level.

Each regulatory corporate disclosure is published in both German and English. The additional publication of corporate disclosures in English enables international investors to react sufficiently fast to new information. Given the regulatory background, the content of both the German and English disclosures should be identical. The corporate disclosures' associated companies are, at least, traded on the fully electronic trading system Xetra. For each security, an (open) limit order book is provided on Xetra (similar to the exemplary order book shown in Fig. 1). Investors post orders into the limit order book and thereby indicate their willingness to trade. The dataset at hand contains this high-frequency (level-2) order book data, i.e., it allows insights into the order book breadth and depth, including the best ten bid and offer limits and the respective order quantities at those limits. The order book data were extracted from *Thomson Reuters Tick History*. Respective order book data were extracted for event dates, i.e., the publication dates of the ad hoc disclosures, and from the previous ten working days. The respective trading volumes and trading phases (e.g., continuous trading, auction) were also extracted from the *Thomson Reuters* access.

5. Data understanding

Before the dataset at hand can be used to forecast the liquidity changes that are caused by ad hoc disclosures, we investigate whether these news have an impact on the level of liquidity at all. Therefore, an event study is conducted that measures the liquidity impact of the event "publication of an ad hoc disclosure". This arrangement also represents an important part of the data understanding phase in the KDD process.

5.1. Methodology

The liquidity measure used in this study is similar to the *Cost of Round Trip* (*CRT*) proposed by Irvine et al. [23]. *CRT* builds upon the information that is contained in the open limit order book (see Fig. 1); i.e., it measures the ex-ante committed liquidity that is available in the market for immediate execution. We compute $CRT(v)_t$ as follows: Given a certain order book situation (and the limits/quantities, respectively) at time t, the (hypothetical) cost of simultaneously buying and selling volume v is calculated. It should be noted that the more liquid the market is, the lower the cost. The cost figure is divided by the executed € volume, i.e., v, to receive the per € cost of a (hypothetical) roundtrip trade.

High values of $CRT(v)_t$ indicate that the market is illiquid and that implicit transaction costs are expected to be high. Any order book activity, such as the submission, cancellation, adjustment or execution of (limit) orders, obviously changes the $CRT(v)_t$. Because we operate with time periods rather than with single points in time, we must calculate an $averageCRT(v)_{T,[t1,t2]}$ for a specific time interval [$t1$, $t2$] on day T. Thereby, the $CRT(v)_t$ values at fixed points in time, i.e., every ten seconds, serve as input.

To investigate the potential impact from the publication of corporate disclosures on firm liquidity levels, we make use of event study methodology. An approach that is similar to the constant-mean-return model is applied [5]. With this approach, we adjust an observed effect by a mean that has been calculated using historical data prior to the event date (see Formula 1). Calculating the abnormal liquidity measure $ACRT(v)_{[t1,t2]}$, we adjust the $averageCRT(v)_{T0,[t1,t2]}$ at event day $T0$ by the previous N days ($N = 10$)[2] $averageCRT(v)_{T,[t1,t2]}$ for the same time period [$t1$, $t2$]. The adjustment is undertaken for the same (intraday) time period [$t1$, $t2$] because the literature suggests that the liquidity levels systematically change during the course of the day [33]. In general, when $ACRT(v)_{[t1,t2]}$ is positive (>0), the liquidity at event day $T0$ is worse compared to historical levels. In contrast, $ACRT(v)_{[t1,t2]}$ being negative (<0) indicates a higher level of liquidity at $T0$ compared to the past.

$$ACRT(v)_{[t1,t2]} = \frac{averageCRT(v)_{T0,[t1,t2]} - \frac{1}{N}\sum_{j=1}^{N} averageCRT(v)_{T0-j,[t1,t2]}}{\frac{1}{N}\sum_{j=1}^{N} averageCRT(v)_{T0-j,[t1,t2]}} \quad (1)$$

Because abnormal liquidity levels are intended to be comparable among companies, these are calculated as relative values. Additionally, to ensure that the volumes v (and consequently the *ACRT* values) are comparable among companies, these were derived from historical trade data. For each event, different trade volume metrics (e.g., median trade size, mean trade size) were calculated for the respective one-month period prior to the event date. It follows that the relative sizes of liquidity shocks should be comparable among companies and attributable to corporate disclosure content (only).

5.2. Event study results

$ACRT(v)_{[t1,t2]}$ is calculated for different volumes v and different time periods [$t1$, $t2$]. If the median of $ACRT(v)_{[t1,t2]}$ turns out to be significantly different from zero (Wilcoxon signed rank test), then the chosen event type *regulatory-driven corporate disclosures* constitutes a critical market event for which significantly higher/lower liquidity levels (and consequently lower/higher transaction costs) can be expected.

Table 1 depicts median $ACRT(v)_{[t1,t2]}$ values for different 15-minute time intervals (whereas 0 is the time of publication) and varying volumes v. The first two inputs *median* and *4*median* depend on each security's previous one-month trade statistics.[3] The last input *25 k*, however, is fixed/equivalent for each security, i.e., volume v = € 25,000. The results in Table 1 provide the following insights:

First, it can be observed that there are no significant abnormal liquidity levels prior to the publication of corporate disclosures, e.g., [−15, 0]. This finding provides evidence that the chosen event type

[2] During the step of *creating a target dataset*, corporate disclosures related to the same company that were published within a time window of the previous $N = 30$ days were not included (see above). However, the time period that was used to calculate the "historical" liquidity levels is $N = 10$ days. The time difference of 20 days allows the liquidity levels to definitely revert to normal levels in the case of a relevant event 31 days prior to the current event.

[3] We take into account the previous one-month trade statistics because publications on trading statistics in nearly all markets in the world are based on a one-month period (see e.g. reports provided by the World Federation of Exchanges http://world-exchanges.org/statistics, the Federation of European Stock Exchanges http://www.fese.eu/en/?inc=page&id=10, or the Thomson Reuters Monthly Market Share Reports http://thomsonreuters.com/monthly-market-share-reports/) and in order to avoid intra-month effects.

Table 2
Median $ACRT(v)_{[t1,t2]}$ sorted by indices.

		$ACRT(v = 4^*median)_{[t1,t2]}$					
Index	n	[−30, −15]	[−15, 0]	[0, 15]	[15, 30]	[30, 45]	[45, 60]
DAX	122	−0.019	−0.015	0.312***	0.108***	0.064***	0.037***
MDAX	136	−0.023	−0.034	0.341***	0.133***	0.077***	0.037*
SDAX	157	−0.065	−0.023	0.183***	0.086***	0.058**	0.037**

***/**/* indicate significance at the 1%/5%/10%-level.

actually contains new and previously unknown information that has not been widely anticipated by market participants. It is therefore expected that forecasting of such liquidity shocks is especially challenging for models that are based solely on historical quantitative data.

Second, we find strong empirical evidence that transaction costs increase subsequent to the publication of corporate disclosures, e.g., [0, 15]. This finding is most likely due to the fact that the disclosures' contents persuade traders to adjust their valuations of the respective company and adjust their existing limit orders in the order book accordingly. During the adjustment process, fewer limit orders (or limit orders with a lower volume) remain in the market, and therefore, the cost of execution increases (i.e., the liquidity decreases).

Third, the results appear to be robust to different volume *v* inputs. Both the sizes and signs of the results (the median values) are – as expected – comparable among inputs, i.e., *median*, *4*median*, and *25 k*. Therefore, the results of the analyses below will be shown only for the input *4*median*.

It can also be observed that there are no systematic differences in the median $ACRT(v = 4^*median)_{[t1,t2]}$ values, if they are calculated for the index subgroups (Table 2). The only small difference can be found with the fact that the levels of significance decrease in case of the medium (MDAX, period [45, 60]) and small capitalization indices (SDAX, periods [30, 45] and [45, 60]).

In line with Graham et al. [15], we suppose that the reaction of the liquidity levels to new information depends on the news types. We therefore construct news/corporate disclosure sub-samples according to the categories proposed by Leis and Nowak [29]. The sub-sample analysis provides the following insights (Table 3):

First, for those categories that have a sufficient number of corporate disclosures (i.e., at least 10 disclosures), we can observe highly significant abnormal liquidity levels subsequent to the publication of the corporate disclosures, e.g., for [0, 15]. Nonetheless, both the size and length of the liquidity impacts appear to vary among news categories. While the liquidity levels revert to *normal* levels, already 15 min after publication of *corporate action* disclosures, the liquidity impact of *financial statement* disclosures and *miscellaneous* disclosures appears to be more persistent. We can therefore conclude that the durability of liquidity impacts depends on the type of news. The observed differences in durability might be the result of continuously varying opinions of market participants. In other words, the content of those corporate disclosures is not unambiguous.

Second, we can observe significant market reactions prior to the publication of *corporate action* disclosures. Negative $ACRT(v)_{[t1,t2]}$ values, however, denote abnormally high liquidity levels (i.e., low implicit transaction costs). Market participants appear to be able to anticipate *corporate action* disclosures. Whether this information is based on insider information or publicly available information and rumours is, however, not observable within our setting.

Given the above results, one might typically assume lower liquidity levels (i.e., higher implicit transaction costs) subsequent to the publication of corporate disclosures. Therefore, the simplest strategy to avoid high implicit transaction costs would be to either execute orders immediately at t_0 (*naive* strategy) or wait for execution until the liquidity reverts to normal levels. The latter case would, however, incur waiting/opportunity costs. Moreover, as shown above, the length of time after which the liquidity levels revert to *normal* levels varies among news types.

With respect to the data understanding phase of our research approach, we conclude that the corporate disclosures at hand do have an impact on the corresponding liquidity levels. As a result, forecasting this liquidity impact appears to be promising because investors as well as automated trading engines must adjust their strategies according to the expected liquidity impact to reduce the implicit transaction costs.

5.3. Document labelling according to the liquidity impact

Supervised learning requires us to label the documents in the document collection according to pre-defined objectives. Our objective is to forecast news-related liquidity impacts. As shown in the course of the event study, we can (on average) observe positive $ACRT(v)_{[t1,t2]}$ values, i.e., higher transaction costs, which occur subsequent to the publication of corporate disclosures. Therefore, our naive response would be to expect lower liquidity levels (i.e., higher implicit transaction costs). If we, however, take a closer look at the distribution of the $ACRT(v)_{[t1,t2]}$ values (Fig. 3), we note that certain corporate disclosures are associated with negative $ACRT(v)_{[t1,t2]}$ values (class *negative*). Given our otherwise naive response, we are especially interested in identifying those corporate disclosures that are associated with negative $ACRT(v)_{[t1,t2]}$ values, i.e., those cases that have implicit transaction costs that are below historical levels and consequently liquidity levels that are above historical levels.

The empirical distribution of the $ACRT(v)_{[t1,t2]}$ values in Fig. 3 shows that the 25%-quartile by coincidence separates those cases quite well from the overall distribution that are of most interest to us, i.e., those with a negative $ACRT(v)_{[t1,t2]}$ value. In other words, we use the 25%-quartile for labelling purposes. Consequently, each corporate disclosure

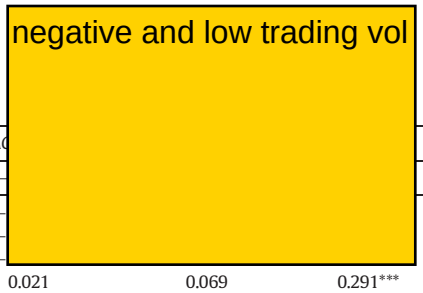

Table 3
Median $ACRT(v)_{[t1,t2]}$ sorted by news categories.

News category	n	[−30, −15]	[−15, 0]	[0, 15]	[15, 30]	[30, 45]	[45, 60]
(1) Financial statement	122				0.119***	0.101***	0.093***
(2) Dividend announcement	39				0.038	−0.013	0.043
(3) Corporate action	37				0.082	0.019	0.017
(4) M&A transaction/reorganization	80	0.021	0.069	0.291***	0.163***	0.070**	−0.025
(5) Personnel	61	−0.030	−0.016	0.076**	0.031	0.034	0.027*
(6) Litigation [a]	5	−0.234	0.137	0.011	0.144	0.407	0.158
(7) Order situation [a]	4	−0.058	−0.083	0.791	0.716	0.244	0.188
(8) Investments [a]	8	0.081	0.023	0.015	0.134	−0.026	−0.127
(9) Miscellaneous	59	0.024	0.029	0.339***	0.093***	0.035**	0.054**

***/**/* indicate significance at the 1%/5%/10%-level.
[a] Please note that the number of observations, i.e. n, is too small to achieve reliable statistically grounded insights for this news category.

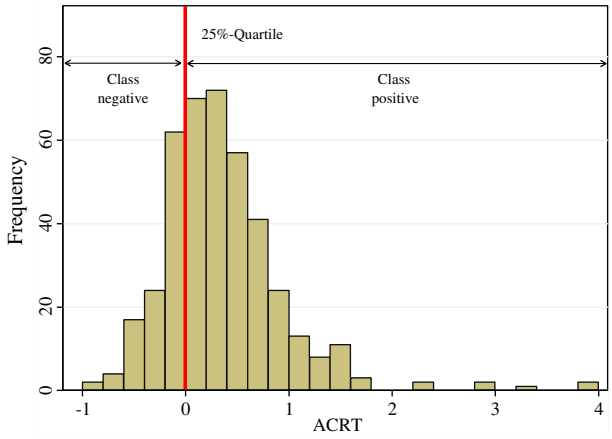

Fig. 3. ACRT($v = 4^{*}median$)$_{[0,15]}$ distribution.

is assigned to the classes *positive* or *negative*, depending on whether their $ACRT(v)_{[t1,t2]}$ is above or below the 25%-quartile of all of the documents' $ACRT(v)_{[t1,t2]}$. The above event study provides evidence that the strongest reaction to the publication of corporate disclosures can be observed during the first 15 min. Consequently, this interval is of most interest to investors, and as a result, the time period [0, 15] is used for labelling.

6. Data cleaning & pre-processing followed by data reduction

At the beginning of the data cleaning & pre-processing phase, a dictionary of words and phrases that adequately describes the document collection is generated. In this way, a simple *StringTokenizer* [45] splits up the whole text into individual units in lower case (i.e., by also applying a *Lower_Case_Converter*). Although we basically follow a bag-of-words approach, in which grammar or word order is not accounted for, we additionally create term *n-grams* ($n = 2$) to grasp the most important word combinations. This approach allows us to better interpret word combinations, such as *not_good* and *not_bad*, which would otherwise have been assessed in isolation.

In addition to *StringTokenizer* units, we also create character *n-grams* of length *n* of each token in a document. In line with Braschler and Ripplinger [3], we make use of 6-grams. Note that both the character *n-grams* and *StringTokenizer* units are part of the final feature set. We also create character *n-grams* because the German language is especially coined by a rich set of possible inflections and word concatenations that might not be properly accounted for by stemmers [3]. Nonetheless, we also map different grammatical forms of a word to a common stem by applying either the *Porter Stemmer* [38] or the *Lovins Stemmer* [32]. The stemmer procedure is applied to the *StringTokenizer* units only.

Table 4
Exemplary pre-processing setup that illustrates the feature number reduction.

Pre–processing steps	No. of features			
	Language: German		Language: English	
	Complete ad–hoc news	Headline only	Complete ad–hoc news	Headline only
Tokenizer	13,038	1,306	8,626	1,383
→ *Lower_Case_Converter*	12,114	1,253	7,216	1,216
→ *Token_Length_Filter* (min. 4)	11,546	1,120	6,682	1,074
→ *Stemmer* (Porter)	10,370	1,075	4,628	904
→ *Pruning* (below 3; above 25)	3,118	233	1,732	277

To eliminate noise, e.g., words that have little meaning but frequent appearance, we apply a threshold on the number of documents that each token occurs in. Pre-tests revealed that useful thresholds for the dataset at hand are 3 for too infrequent words and 25 for too frequent words. Words that are above and below those thresholds are not included in the feature set. In addition, tokens that do not fulfil the minimum length requirement of 4 were removed from the feature set (*Token_Length_Filter*). An additional *stop word* list is not made use of, to ensure comparability for the language comparison, i.e., German/English.

Moreover, the *Chi-Squared* metric serves as an additional method for data reduction. Thereby, the feature set is further reduced by scoring (filtering) the features according to their *Chi-Squared* weightings. Our pre-tests revealed that filtering features according to their *Chi-Squared* weightings provides better results than filtering by *Information Gain*. Those *k* features that have the highest weightings remain in the feature set.

Optimal values of *k* are derived within a simple grid-search loop iteration. In this way, optimal parameter values are obtained by varying combinations of assigned parameter values (i.e., between $k = 50$ and $k =$ the total number of tokens). Whenever optimal parameter values were derived by a grid-search, it is marked as such in the results section.

At the end of the process, each document is represented by the previously extracted and selected number of features. The respective feature weightings in the document-feature matrix *W* is given by *tf.idf* [30].

Table 4 exemplarily depicts how different pre-processing steps reduce the feature set. The insights from this illustration are as follows:

First, the feature set that results from the complete corporate disclosures (*news*) is, of course, larger than the feature set that results from the headlines only (*headline*). As already suggested above, the English feature set is much smaller than the German feature set [3]. After having applied different pre-processing steps, however, the *headline* feature sets are similar in size.

Second, already the conversion of all of the tokens (words) to lower cases has an effect on the size of the feature sets. The largest reduction of features is, however, achieved by pruning.

7. Selection of data mining method and data mining algorithm

7.1. Classification technique

Because the aim of our study is to investigate whether unstructured data can be used for liquidity forecasts after the publication of ad hoc disclosures, we concentrate on the application of a single machine learning technique instead of comparing different techniques, as is proposed by the KDD process of Fayyad et al. [11]. For selecting an appropriate machine learning technique, we consider comparative empirical studies. These studies provide evidence that the classification performance of *SVM* is superior to other data mining techniques (e.g., [24]). In addition, *SVM* "is usually less vulnerable to the overfitting problem" and "the solution of SVM is always unique and globally optimal" [21]. As a result, we make use of *SVM* within our text mining approach.

SVM was first introduced by Vapnik [43] for solving two-class recognition problems. The basic idea is to find a decision surface that maximises the margin between the data points, i.e., the classes, by means of structural risk minimisation. In a case in which there are originally non-separable data points, the original data vectors can be mapped to a higher dimensional space to achieve linear separability. To reduce the complexity, *kernels*, i.e., functions in lower dimensional space that exhibit similar behaviour as the original functions in higher dimensional space, are applied. We make use of a linear *kernel* similar to Hsu et al. [20], who provide evidence that a linear *kernel* appears to be sufficient whenever the number of features is exceptionally large.

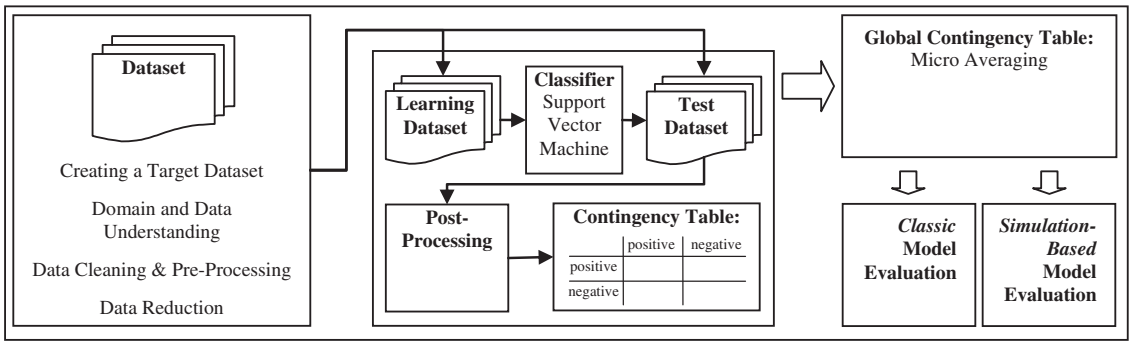

Fig. 4. Intraday text mining approach setup.

When applying *SVM* with a linear *kernel*, primarily the cost parameter C must be optimised. Thus, we also apply the above-mentioned grid-search loop optimisation approach. In line with Hsu et al. [20], we use exponentially growing sequences of parameter C as input into grid-search optimisation, i.e., $C = 2^{-5}, 2^{-3},..., 2^{15}$.

7.2. Post-processing

SVM delivers confidence values for each class, to assure that the prediction is actually *true positive* [34]. Corporate disclosures are assigned to the classes *positive* or *negative*, depending on whether the confidence value is above or below a certain (learned) threshold. The variation in the thresholds can be applied as a post-processing step to account for imbalanced datasets or unequal classification costs [47]. We are confronted with both: First, the class *negative*, which is of most interest to us, contains only 25% of all corporate disclosures. Second, the cost for falsely classifying *negative* documents as *positive* is higher than falsely classifying *positive* documents as *negative*. This relationship occurs because our objective is to precisely identify those corporate disclosures that are associated with abnormally low transaction costs after publication, i.e., class *negative*. To overcome these problems, we conduct cost-sensitive learning implemented as a post-processing step [22]. Thereby, a *ThresholdFinder* [34] uses the confidence values to turn the *SVM* into a cost-sensitive learner.

8. Interpretation of results: *Classic* model evaluation

To be able to properly interpret the performance of the proposed text mining setup, this section presents a *classic* evaluation by means of common machine learning performance metrics.

8.1. Classic model evaluation setup

As shown in Fig. 4, the whole dataset is split up into a learning dataset and a test dataset to ensure that model evaluation is independent from model building. Because the dataset is comparatively small, we do not conduct a one-time split; instead, we follow an *m-fold* ($m = 10$) cross-validation approach [44].

Each test sub-sample contingency table is aggregated to create a global contingency table (micro averaging) (Table 5). The global contingency table is used to calculate the *classic* performance measures of *accuracy* [(a + d) / n; n = a + b + c + d], *recall* [class *positive*: a / (a + c)], and *precision* [class *positive*: a / (a + b)] [19]. In this way, *accuracy* denotes the percentage of examples that are classified correctly, *recall* represents the percentage of positive examples that are classified as positive, and *precision* indicates the percentage of examples that are correctly predicted as positive.

Accounting for the inherent trade-off between *precision* and *recall*, we additionally calculate the *F*1 measure by van Rijsbergen [42], where *recall* and *precision* are given equal weight (Eq. (2)).

$$F_1 = (2 \cdot recall \cdot precision)/(recall + precision) \quad (2)$$

8.2. Classic model evaluation results

Classification results for complete corporate disclosures (*news*) and headlines only (*headline*) in both German and English provide the following general insights (see Table 6): High (misclassification) costs of the class *negative*, e.g., 0.9, result in a high *precision* figure and a low *recall* figure. In other words, there are only very few corporate disclosures that are assigned to the class *negative*, but those that are assigned truly belong to this class. This finding is due to the fact that corporate disclosures are assigned to the class *negative* only if the respective *SVM* confidence value is quite high, i.e., the classifier is *certain* that the disclosure actually belongs to the respective class.

The *precision* figure of 100% (~90%) for $SVM_{(0.1;0.9)}$ provides strong evidence that the proposed text mining approach can precisely identify some of those corporate disclosures whose associated securities' liquidity levels are not negatively affected by the disclosure contents. The high *precision* figure comes at the cost of low *recall*: We merely capture a small number of relevant class *negative* corporate disclosures. Therefore, one might be willing to accept a certain number of *false positives* to increase the number of caught class *negative* corporate disclosures, e.g., $SVM_{(0.5;0.9)}$. In the end, it is up to the developers of *automated trading engines* to decide on the classified documents' confidence values. For example, a *safe* recommendation in the form of a trading signal would be provided at a *precision* of 100% with $SVM_{(0.1;0.9)}$.

Accuracy figures are slightly larger than the 75% all-*positive* benchmark would suggest. However, because we are primarily interested in precisely identifying class *negative* corporate disclosures, the benchmark is of little use. We merely show *accuracy* figures for reasons of completeness.

In terms of distinctions between languages (German vs. English) or completeness of texts (complete news vs. headlines only), we cannot find consistent systematic differences. However, utilising complete German texts instead of complete English texts leads to slightly better results.

Table 5
Exemplary illustration of the (global) contingency table.

	True *positive*	True *negative*
Predicted *positive*	a	b
Predicted *negative*	c	d

In terms of the (pre-processing) configuration, it can be observed that many *optimal* "complete news" result set combinations contain term n-gram tokens, whereas many "headline only" result set combinations contain character n-gram tokens. Therefore, for both German and English texts, n-grams should ideally be part of the generated feature sets. Moreover, pruning (thresholds) on the number of documents that each token occurs in rarely leads to superior results.

To summarise, the proposed system can produce a trading signal that indicates whether the underlying regulatory corporate disclosures will (most likely) cause abnormally high or low transaction cost levels during the 15 min subsequent to their publication. Thereby, the developer of an automated trading engine must decide whether high precision (which results in low recall) or high recall (which results in low precision) is the primary goal. Furthermore, this proposed trading signal can serve as an additional input to existing trading models that are purely based on quantitative data.

9. Acting on the discovered knowledge: *Simulation-based* model evaluation

In addition to the above *classic* model evaluation, we introduce a novel domain-specific *simulation-based* evaluation approach that aims at acting on the discovered knowledge. In general, a *simulation-based* model evaluation allows for additional statistical analysis and provides insights into the results' robustness.

Previous applications of text mining techniques in the financial industry have highlighted the need for domain-specific evaluation metrics [16]. Related research has focused on forecasting stock prices or stock price volatilities, whereas these studies evaluate the proposed approaches by performing investment simulations, i.e., a stock is assumed to be bought (sold) when the price is predicted to rise (fall), and the returns achieved with these strategies are used to evaluate the performance [36,40]. However, such a simulation approach is not appropriate for evaluating the implicit transaction costs because it ignores the timing of the orders. Thus, a new simulation setup must be developed. The proposed simulation setup constitutes an *automated trading engine* use case.

Moreover, this simulation allows us to quantify the economic value of the text mining system. *Classic* model evaluation results provide insights merely into whether certain events have been classified correctly. However, *classic* model evaluation does not allow drawing conclusions about the economic value, i.e., whether the "right" classification is actually of any use.

9.1. Simulation-based model evaluation setup

Within the simulation, an *automated trader* receives the order to execute volume V during the time interval M [$m1$, $m2$]. The goal is to minimise the implicit transaction costs. To achieve this goal, an optimal execution strategy must be determined. Within this context, we account for the results of our m-fold cross validation setup (see Fig. 5). Each corporate disclosure in the m test datasets is classified by the above proposed text mining approach.

If the corporate disclosure is classified as belonging to the class *positive*, then we expect the liquidity of the underlying security to decrease during the time interval [0, 15]. Consequently, the *automated trader* should prefer to execute at the very beginning of the interval, i.e., immediately after the release of the news but prior to the liquidity impact or at the end of the interval. Because the above event study analysis has provided evidence that the length of time after which the liquidity levels revert to *normal* levels varies, we decided in favour of the first approach, i.e., execution at the very beginning (*naive strategy*). Thus, the *automated trader* performs analysis on the incoming disclosure immediately, i.e., within milliseconds and, therefore, very close to $t + 0$, places an appropriate order. In this way, the *automated trader* can make use of the available (high) level of liquidity *before* other traders (human traders) change their bids and offers in response to the news disclosure.

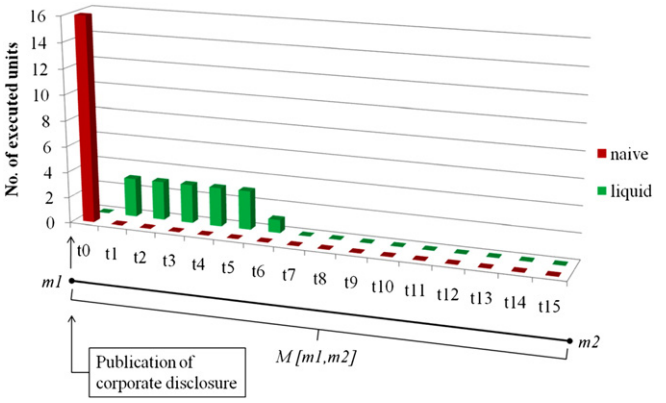

Fig. 6. Illustration of the applied trading strategies naive and liquid.

In contrast, if the corporate disclosure is classified as belonging to the class *negative*, then we expect the liquidity of the underlying security to increase – or at least not to decrease – during the time interval. To take advantage of the liquidity increase, we introduce a *liquid strategy*. Therefore, the volume V is split into different orders that are executed step-by-step.

The execution strategies explained above are also illustrated in Fig. 6. Please note that the *automated trader* should follow a strategy that simultaneously profits from advantageous liquidity levels and also accounts for the market impact costs from large trades. For example, we expect the volume of the first naive strategy to be executed at still *normal* liquidity levels (t_0). Nonetheless, we also expect a large market impact that is associated with the executed volume.

In terms of modelling, the market impact is assumed to be merely temporary, i.e., it does not last until the next execution time t_i.

Furthermore, we specify the following model assumptions: The choice among alternative trading strategies is *static* in our simulation model, i.e., the trading strategy chosen at $m1$ cannot dynamically be altered during M. Moreover, we do not explicitly model buy or sell orders. Instead, each time that the *automated trader* triggers an execution of size v_i, a cost measure similar to the above-defined $CRT(v_i)_{ti}$ is used as a proxy for the incurred transaction costs. Please note that the cost measure used here is not an abnormal measure, i.e. not *ACRT*. The

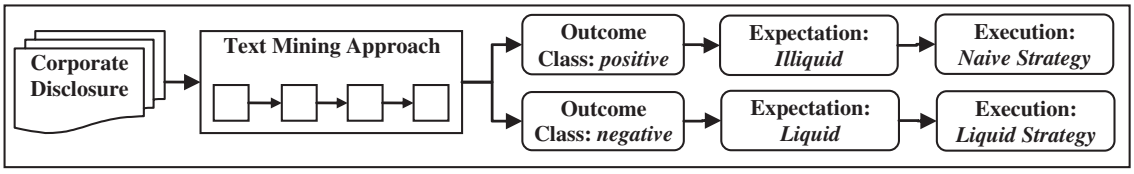

Fig. 5. Setup of *simulation-based* model evaluation.

Table 6
Classic model evaluation results.

Misclassification cost for class (pos./neg.)	Complete ad-hoc news							
	Language: German				Language: English			
	Accuracy	*Precision*	*Recall*	*F1*	*Accuracy*	*Precision*	*Recall*	*F1*
(0.1/0.9)	80.48	100.00	22.12	36.23	80.00	92.00	22.12	35.66
(0.3/0.9)	81.69	72.58	43.27	54.22	80.72	72.22	37.50	49.37
(0.5/0.9)	81.69	70.00	47.12	56.33	80.24	65.71	44.23	52.87
(0.9/0.9)	79.04	57.80	60.58	59.16	75.66	51.33	55.77	53.46
	Headline only							
(0.1/0.9)	78.07	93.33	13.46	23.53	79.76	91.67	21.15	34.37
(0.3/0.9)	79.26	72.50	27.88	40.27	80.24	80.56	27.88	41.42
(0.5/0.9)	79.04	68.89	29.81	41.61	80.00	78.38	27.88	41.13
(0.9/0.9)	77.83	58.82	38.46	46.51	77.11	55.70	42.31	48.09

Complete news/German: (0.1/0.9) C = 2, k = 500, Porter stemmer, term n-gram, pruning = yes//(0.3/0.9) C = 2048, k = 500, Lovins stemmer, term n-gram, pruning = no//(0.5/0.9) C = 32, k = 200, Lovins stemmer, term n-gram, pruning = yes//(0.9/0.9) C = 32768, k = 50, Porter stemmer, character n-gram, pruning = no. *Complete news/English*: (0.1/0.9) C = 32, k = 200, Porter stemmer, no n-gram, pruning = no//(0.3/0.9) C = 0.03125, k = 200, Lovins stemmer, term n-gram, pruning = no//(0.5/0.9) C = 0.5, k = 500, Lovins stemmer, term n-gram, pruning = no//(0.9/0.9) C = 2048, k = 200, Porter stemmer, term n-gram, pruning = no. *Headline only/German*: (0.1/0.9) C = 2048, k = 50, Porter stemmer, character n-gram, pruning = yes//(0.3/0.9) C = 0.5, k = 50, Porter stemmer, character n-gram, pruning = no//(0.5/0.9) C = 0.5, k = 50, Porter stemmer, character n-gram, pruning = no//(0.9/0.9) C = 32, k = 50, Porter stemmer, character n-gram, pruning = no. *Headline only/English*: (0.1/0.9) C = 2048, k = 200, Lovins stemmer, character n-gram, pruning = no//(0.3/0.9) C = 128, k = 50, Lovins stemmer, character n-gram, pruning = no//(0.5/0.9) C = 2048, k = 50, Lovins stemmer, character n-gram, pruning = no//(0.9/0.9) C = 8, k = 200, Lovins stemmer, term n-gram, pruning = no.

transaction costs for each *strategy* and event *j* are calculated according to Formula 3:

$$Cost_{Strategy,j} = \sum_{i=0}^{15} CRT(v_i)_{t_i}. \quad (3)$$

The simulation time interval *M* is equivalent to the period that is used for labelling, i.e. [0, 15]. Equivalent to the above approach, the volume *v* is derived from the trade statistics, i.e., it is *4*median* in this case, and it must be executed during this time interval for each security that is subject to a corporate disclosure. Executions of size v_i are made at fixed one-minute intervals ($t_0, t_1, t_2, \ldots, t_{15}$). In this way, t_0 constitutes the corporate disclosure publication time at the publication day. It is assumed that *automated traders* – even with a simple trading signal as *news* or *no news* – can react to the publication of corporate disclosures within milliseconds, i.e., while the liquidity is still at *normal* levels (at t_0). To further rule out distortions because of systematic differences between securities, we calculate a cost ratio (Formula 4). The cost ratio directly compares the costs that are associated with each strategy for each event.

$$R_{Cost} = \frac{Cost_{naive,j}}{Cost_{liquid,j}} \quad (4)$$

Having already provided evidence that the liquidity levels are expected to decrease for the majority of events (see Fig. 3) and that this effect is larger in size than the market impact effect of large trades, the costs that are associated with the *naive strategy* ($Cost_{naive,j}$) should be lower than the costs that are associated with the *liquid strategy* ($Cost_{liquid,j}$). Consequently, we expect the cost ratio R_{Cost} to be lower than one for all (and *positive*) events. If our text mining approach can precisely identify *negatively* labelled events, then the cost ratio should be larger than one for the (*negative*) events.

News assigned to class *positive* : $Cost_{naive} < Cost_{liquid}; R_{Cost} < 1$
News assigned to class *negative* : $Cost_{naive} > Cost_{liquid}; R_{Cost} > 1$

To further statistically explore this assumption, the corresponding null and alternative hypotheses are formulated and tested:

$$H_0 : \mu(R_{Cost}) = 1 \quad \text{vs.} \quad H_A : \mu(R_{Cost}) \neq 1.$$

Hereby, μ constitutes the *median*, and the hypotheses are tested by means of a Wilcoxon signed rank test.

Table 7
Descriptive and test results for R_{Cost} (*naive/liquid*).

Misclassification cost for class (pos./neg.)	Result subset actually classified as	*Median* R_{Cost}			
		Language: German		Language: English	
		Complete ad-hoc news	Headline only	Complete ad-hoc news	Headline only
	All news	0.95			
(0.1/0.9)	*Positive*	0.94	0.94	0.94	0.93*
	Negative	1.14*	1.32*	1.01	1.20***
(0.3/0.9)	*Positive*	0.92**	0.93	0.93*	0.94*
	Negative	1.12***	1.01	1.12	1.16*
(0.5/0.9)	*Positive*	0.90**	0.93	0.91**	0.94
	Negative	1.17***	1.01	1.11**	1.14*
(0.9/0.9)	*Positive*	0.88***	0.94	0.88**	0.91*
	Negative	1.09**	0.99	1.09**	1.05*

***/**/* indicate significance at the 1%/5%/10%-level.

9.2. Simulation-based model evaluation results

Descriptive and test results for the *simulation-based* model evaluation and varying *SVM* misclassification cost inputs provide the following insights (see Table 7).

First, the median cost ratio for all of the events (n = 415) is below 1. In other words, for all of the events, the *naive strategy* is associated with lower costs than the *liquid strategy*. Nonetheless, it shall be noted that the null hypothesis is not rejected. Therefore, the figure should serve to obtain first insights that concern the relationship between the two strategies.

Second, the *negative* subclass is often associated with statistically significant median R_{Cost} values that are above 1. At the same time, the respective *positive* subclass median R_{Cost} values are below 1.

Third, complete input texts in German appear to provide slightly better results than complete input texts in English. The results differ when only headlines are taken into account: subclass *negative* R_{Cost} values are significant and above 1 for "English headlines only", whereas these are statistically significant and above 1 just in one configuration for "German headlines only". However, the configuration (0.1/0.9) still leads to a higher median cost ratio and thus better execution performance compared to the English language.

To summarise, if the trading signal that is produced by the proposed text mining approach is followed, then the liquidity levels can be

forecasted correctly in order to decrease the implicit transaction costs. As Table 7 shows, for negatively classified complete German corporate disclosures, R_{Cost} is significantly above one. In the case of complete German ad hoc disclosures, a trader who follows the naive approach and executes the whole volume at t_0 (instead of following the proposed liquid strategy) would have to bear implicit transaction costs that are 9%–17% higher compared to the recommendation of our proposed text mining system. Similar results are also found for English corporate disclosures. These findings provide evidence that our proposed text mining system works well and that the proposed approach can be seen as economically relevant.

10. Conclusion

Text mining techniques are already applied in various research projects and practical applications to electronically classify financial news and/or to forecast price changes in securities markets. However, most research is focused on the prediction of future price changes of a security. Given that liquidity constitutes one of the most important determinants of (implicit) transaction costs, we aimed to investigate whether text mining allows us to predict future levels of liquidity.

We follow the KDD process proposed by Fayyad et al. [11]. However, given the specific environment of the financial domain, it is necessary to adjust and "shape" the KDD process to fit domain-specific requirements. Based on an in-depth understanding of the respective industry, we propose such an adapted KDD process.

First, we conduct an event study, which is a common approach in the financial domain, to understand the data. The event study provides empirical evidence that the publication of regulatory corporate disclosures is followed by abnormal liquidity levels. This finding is consistent with existing beliefs about how limit order traders update their orders upon the arrival of new information. We, however, do not find consistent evidence of abnormal liquidity levels prior to the publication of corporate disclosures. It follows that *automated traders* should ideally include information on the publication of corporate disclosures into their models.

Second, we develop a domain-specific *simulation-based* evaluation approach to assess the economic value that is added by the proposed text mining system. Both *classic* and *simulation-based* model evaluation results provide evidence that the trading signal indicates some of those corporate disclosures, entailing the lowest expected future transaction costs. Consequently, we have shown that it is of economic value to adhere to the proposed text mining approach, i.e., to include information on the publication of corporate disclosures into *automated traders'* models. Moreover, having implemented and tested an IT artefact, we have also shown how such a text mining approach might look.

To summarise, following a structured domain-specific KDD process, it is possible to extract useful information from unstructured qualitative data to predict future levels of liquidity. In terms of a text mining setup, we found weak evidence that the German language is more suitable than the English language in case of complete disclosures taken into account. One possible explanation might be the fact that (German) concatenations are useful within a *bag of words* approach. As a practical implication, international investors may benefit from localising their text mining systems, at least within German-speaking countries.

Given the above results, this paper contributes in terms of both methodology (e.g., the adapted KDD process, event study, simulation setup) and practical relevance (e.g., a detailed description of the IT artefact). Being a highly relevant group of traders and despite their technical capabilities, *automated traders* require an appropriate decision support system as well. In this research project, we proposed and successfully tested different ways of enhancing *automated trading engines* to address news-related liquidity shocks in a timely manner. Future work will concentrate mainly on solving the current limitations of this research, i.e., the proposed forecasting approach shall be compared to existing quantitative forecasting approaches. Because this paper's proposed IT artefact is not intended to replace existing systems (and is instead intended to complement them), future work will therefore concentrate on the integration of our trading signal into existing execution models.

References

[1] J. Bikker, L. Spierdijk, R. Hoevenaars, P.J. van der Sluis, Forecasting market impact costs and identifying expensive trades, Journal of Forecasting 27 (1) (2006) 21–39.
[2] N. Bissantz, J. Hagedorn, Data mining, business & information systems engineering 1 (1) (2009) 118–122.
[3] M. Braschler, B. Ripplinger, How effective is stemming and decompounding for German text retrieval, Information Retrieval 7 (3) (2004) 291–316.
[4] R. Brooks, A. Patel, T. Su, How the equity market responds to unanticipated events, Journal of Business 76 (1) (2003) 109–133.
[5] J.Y. Campbell, A.W. Lo, A.C. MacKinlay, The Econometrics of Financial Markets, Princeton University Press, Princeton, N.J., 1997
[6] R. Coggins, M. Lim, K. Lo, Algorithmic trade execution and market impact, IWIF Working Paper, University of Sydney, 2006.
[7] Deutsche Börse AG, Integrity for financial markets, Annual Report 2008, 2008. (http://deutsche-boerse.com/dbag/dispatch/en/binary/gdb_content_pool/imported_files/public_files/10_downloads/12_db_annual_reports/2008/GB_komplett_2008.pdf (2012-03-06)).
[8] V. Dhar, R. Stein, Intelligent decision support methods, The Science of Knowledge Work, Prentice Hall, Upper Saddle River, NJ, 1997.
[9] I. Domowitz, J. Glen, A. Madhavan, Liquidity, volatility and equity trading costs across countries and over time, International Finance 4 (2) (2001) 221–255.
[10] I. Domowitz, H. Yegerman, The cost of algorithmic trading — a first look at comparative performance, in: B. Bruce (Ed.), Algorithmic Trading: Precision, Control, Execution, Institutional Investor Inc., New York, USA, 2005, pp. 30–40.
[11] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Magazine 17 (3) (1996) 37–54.
[12] T. Foucault, M. Pagano, A. Röell, Market liquidity, Theory, Evidence, and Policy, Oxford University Press, Oxford, UK, 2013.
[13] P. Gomber, M. Gsell, Catching up with technology — the impact of regulatory changes on ECNs/MTFs, Competition and Regulation in Network Industries 1 (4) (2006) 535–557.
[14] P. Gomber, U. Schweickert, E. Theissen, Zooming in on liquidity, 31st Annual Meeting of the European Finance Association, Maastricht, Netherlands, 2004, pp. 1–34.
[15] J. Graham, J. Koski, U. Loewenstein, Information flow and liquidity around anticipated and unanticipated dividend announcements, Journal of Business 79 (5) (2006) 2301–2336.
[16] S.S. Groth, J. Muntermann, Supporting investment management processes with machine learning techniques, Proceedings of the 9th Internationale Tagung Wirtschaftsinformatik, vol. 2, Österreichische Computer Gesellschaft, Vienna, Austria, 2009, pp. 275–284.
[17] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems 50 (4) (2011) 680–691.
[18] J. Han, M. Kamber, Data mining, Concepts and Techniques, 2nd ed., Elsevier; Morgan Kaufmann, San Francisco, 2006.
[19] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, GLDV Journal for Computational Linguistics 20 (1) (2005) 19–62.
[20] C.W. Hsu, C.C. Chang, C.J. Lin, A Practical Guide to Support Vector Classification, National Taiwan University, 2003. (http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (2011-10-16)).
[21] W. Huang, Y. Nakamori, S. Wang, Forecasting stock market movement direction with support vector machine, Computers and Operations Research 32 (2005) 2513–2522.
[22] M. Ikonomakis, S. Kotsiantis, V. Tampakas, Text classification using machine learning techniques, WSEAS Transactions on Computers 4 (8) (2005) 966–974.
[23] P. Irvine, G. Benston, E. Kandel, Liquidity beyond the inside spread: measuring and using information in the limit order book, Working Paper, Emory & Hebrew University, 2000.
[24] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 137–142.
[25] D.B. Keim, A. Madhavan, The cost of institutional equity trades, Financial Analysts Journal 54 (4) (1998) 50–69.
[26] I. Krinsky, J. Lee, Earnings announcements and the components of the bid-ask spread, Journal of Finance 51 (4) (1996) 1523–1535.
[27] L.A. Kurgan, P. Musilek, A survey of knowledge discovery and data mining process models, The Knowledge Engineering Review 21 (1) (2006) 1–24.
[28] C.M.C. Lee, B. Mucklow, M. Ready, Spreads, depths, and the impact of earnings information, Review of Financial Studies 6 (2) (1993) 345–374.
[29] J. Leis, E. Nowak, Ad-hoc-Publizität nach § 15 WpHG, Schäffer-Poeschel, Stuttgart, 2001.
[30] D. Lewis, Representation and Learning in Information Retrieval, Dissertation University of Massachusetts, 1992.
[31] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, The Journal of Finance 66 (1) (2011) 35–65.
[32] J. Lovins, Development of a stemming algorithm, Mechanical Translation and Computational Linguistics 11 (1–2) (1968) 22–31.
[33] T. McInish, R. Wood, An analysis of intraday patterns in bid/ask spreads for NYSE stocks, Journal of Finance 47 (2) (1992) 753–764.
[34] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, YALE: rapid prototyping for complex data mining tasks, Proceedings of the ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 2006, pp. 935–940.
[35] M.-A. Mittermayer, G. Knolmayer, Text mining systems for market response to news: a survey, Institute of Information Systems, Working Paper No. 184, University of Bern, 2006.
[36] M.-A. Mittermayer, Forecasting intraday stock price trends with text mining techniques, Proceedings of the 37th Hawaii International Conference on System Sciences, Big Island, Hawaii, USA, 2004.
[37] J. Muntermann, A. Guettler, Intraday stock price effects of ad hoc disclosures: the German case, Journal of International Financial Markets Institutions and Money 17 (1) (2007) 1–24.
[38] M. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 211–218.
[39] A. Ranaldo, Intraday market dynamics around public information arrivals, in: F. Lhabitant, G. Gregoriou (Eds.), Stock Market Liquidity: Implications for Market Microstructure and Asset Pricing, John Wiley & Sons, Hoboken, N.J., USA, 2008, pp. 199–226.
[40] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Transactions on Information Systems 27 (2) (2009) 1–19.
[41] R. Schwartz, R. Francioni, Equity Markets in Action: The Fundamentals of Liquidity, Market Structure & Trading, John Wiley & Sons, Hoboken, N.J., USA, 2004.
[42] C.J. van Rijsbergen, Information Retrieval, 2nd ed. Butterworths, London, 1979.
[43] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, USA, 1995.
[44] I.H. Witten, E. Frank, M.A. Hall, Data mining, Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann Publishers, Burlington, Mass, 2011..
[45] M. Wurst, I. Mierswa, The Word Vector Tool: User Guide, Operator Reference, Developer Tutorial, 2009. (http://heanet.dl.sourceforge.net/project/rapidminer/2.%20Text%20Plugin/4.4/rapidminer-text-4.4-tutorial.pdf (2012-03-06)).
[46] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, W. Lam, Daily stock market forecast from textual web data, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, USA, 1998.
[47] Y. Yang, A study of thresholding strategies for text categorization, Proceedings of the 24th ACM Int. Conference on Research & Development in Information Retrieval, 2001, pp. 137–145.

Sven S. Groth is currently working as a personal advisor to the chief executive officer of an asset management company. Prior to that he worked as a graduate researcher at the E-Finance Lab, an industry-academic partnership between Goethe University Frankfurt am Main, Germany and several (financial) industry partners. During this time he also wrote his PhD thesis titled "automation in securities trading: text mining and algorithmic trading". Sven studied business administration at the European Business School (ebs), Germany and real estate investment at the University of Reading Business School, UK. His research interests include decision support systems and algorithmic trading.

Michael Siering is a graduate researcher at the E-Finance Lab, Frankfurt, Germany. He received his M.Sc. degree in Management from University of Frankfurt and is currently writing his PhD thesis at the Chair of e-Finance. His research interests include financial decision support, financial market surveillance and sentiment analysis.

Peter Gomber holds the Chair of e-Finance at the Faculty of Economics and Business Administration, University of Frankfurt. He is Co-Chair and Member of the Board of the E-Finance Lab. His academic work focuses on market microstructure and auction theory, institutional trading, innovative concepts/technologies for electronic trading and post trading systems and information systems in Finance. Peter Gomber graduated in business administration at the University of Gießen and acquired his PhD at the Institute of Information Systems. He published several articles on the above mentioned topics in international journals and was awarded with the Reuters Innovation Award 2000, the University Award of DAI (Deutsches Aktieninstitut) 1999, the IBM SUR Grant 2007 and several Best Paper Awards at international research conferences.