EPJ.org
EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Entropy-based text feature engineering approach for forecasting financial liquidity changes

Aleksei Riabykh[1], Ilias Suleimanov[2], Ilya Nagovitcyn[2], Denis Surzhko[2], Maxim Konovalikhin[2] and Olessia Koltsova[1*]

[*]Correspondence: ekoltsova@hse.ru
[1]Laboratory for Social and Cognitive Informatics, National Research University Higher School of Economics, 55/2 Sedova St., St. Petersburg, Russia
Full list of author information is available at the end of the article

Abstract

Changes in individual and institutional financial behavior leading to shifts in liquidity flows often depend on events reflected in news. However, the task of establishing relationship between financial behavior and news remains challenging and understudied. We propose a news-based feature generation approach that allows accounting for news events in liquidity flow time-series predicting tasks, thereby improving the forecasting quality. These features are constructed as different types of entropies and calculated at different levels of text abstraction based on word counts, TF-IDF values, probabilistic topics, and contextual embeddings. We show that this feature engineering procedure is effective for predicting changes in two types of liquidity flows: stock market trading volume and the volume of ATM cash withdrawals. As the first type, we use our original collection of 651, 208 business news articles from a Russian news agency dating to 2013-2021 to predict abnormal jumps in the trade volume of 32 leading Russian companies. With our approach, 97% of them experience an increase in the quality of predicting the differences in daily trading volumes from their median values. For the ATM withdrawals task, we test the impact of economic news from three leading Russian media sources (N = 55, 712) on withdrawals from 100 ATMs located in Moscow. For 95% of them we improve the quality of prediction of year-to-year weekly withdrawal volume change. Additionally, we find that some news sources have a higher predictive power than others. The approach is potentially generalizable for other domains of financial behavior across the globe.

Keywords: Feature engineering; Financial time series; Natural language processing; Economic news; Entropy; Stock trade volumes; ATM cash withdrawals

1 Introduction

Liquidity flow forecasting is a paramount task in risk management for banks and other financial institutions. Among other things, it helps them meet liquidity requirements which in turn is essential for the stable operation of financial organizations and for their ability to fulfill their obligations. As a result, liquidity forecasting has developed into a broad field of study [1–3]. Recently, this field has experienced an extensive growth of successful applications of machine learning (ML) for such tasks as the prediction of liquidity volumes and

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Springer

its risk factors [4, 5]. In particular, it has been shown that ML with unstructured text data can provide advantage for a wide range of specific problems, from risk assessment [6] to mobile banking service quality [7]. However, the development of more universal models, applicable to various areas of financial liquidity and able to combine different data sources, is lagging back. Thus, effective preprocessing of heterogeneous data and respective optimal feature engineering has proved a difficult task.

This paper seeks to contribute to the coverage of this gap by proposing a new approach to feature formation for the models predicting changes in liquidity volume. Specifically, we transform unstructured text data from business news into features calculated as entropies of several types and at different levels of abstraction (from word counts to contextual embeddings). We test the efficiency of these features in two predictive tasks, namely in forecasting abnormal spikes in stock exchange trade volume and weekly year-to-year changes in the volume of ATM withdrawals. These two domains usually account for the lion's share of liquidity flows in national economies and are, therefore, of substantial importance.

Thus, stock market trading volume forecasting allows finding insights into the behavior of market participants, developing risk monitoring systems, and enriching algorithmic trading strategies. In this sphere, seemingly unpredictable spikes of trading activity frequently occur as a reaction to a particular media event (e.g., a political or economic shock, or a natural disaster), that affects trading behavior by being covered in news. This explains the helpfulness of news data in the respective predictive models. Likewise, changes in ATM withdrawals are related to various events via news: while periods of economic growth are accompanied by gradually increased cash withdrawals, periods of resonant "black swans" and structural shocks in the economy are accompanied by sharp peaks in withdrawal behavior. Predicting cash flows allows banks to manage their ATM networks better, improve user experience, and minimize network operating costs. In both areas, the usage of news not only increases prediction quality but also contributes to more interpretable results.

This paper aims to show that although stock exchange liquidity and ATM cash liquidity are pretty different, similar news factors can influence liquidity flows both on the stock exchange and ATMs, and, therefore, similar feature engineering works well for both tasks. To do so, we use two unique datasets from Russia collected by us in the period when the Russian financial system was an integral part of the global financial system and thus functioned according to similar principles. In particular, we analyze the data from business news produced by the leading Russian news agencies, trade volume data from the leading Russian companies, and withdrawals from ATMs of a leading Russian bank.

Thus, this paper makes several contributions to the existing research. We are the first to use news as features in the task of ATM withdrawal forecasting, where we achieve a high prediction quality. Moreover, we show that features extracted from news texts and calculated as parameterized and non-parameterized entropies are also able to improve the existing models of stock market trade volume change prediction. Importantly, we treat both tasks as prediction of volume change, not volume per se, which has been rarely done, but which is more relevant for a number of practical tasks of liquidity management. In terms of ML theory development, we demonstrate that entropy-based features work well in the prediction of liquidity in two domains, despite large differences between the latter. This makes the entropy approach promising as a potentially universal approach for feature engineering in many other liquidity domains. Finally, by studying the Russian market, we are able to rank Russian media by the usefulness of the information they carry for financial decision-making.

The remainder of this paper is organized as follows. The Related Work section reviews research on similar topics and their relevance to our task. The Methodology section covers technical aspects of the study and gives definitions used throughout the article. Datasets and Target Variables section describes our datasets, target variables, and methods of preprocessing used in the current study. The Experiments section describes the procedure for constructing liquidity forecasting models and the feature space used. The Discussion of Numerical Results section introduces the results of our experiments and their interpretations and is followed by Conclusion. Appendix A contains tables with raw experimental data. Visualizations of data and models can be found in Appendix B.

2 Related work

Liquidity has been the focus of research for quite a while. Earlier papers model such aspects of liquidity as market structure [8], its influence on a firm's capacity to raise external finance [9], and constraints imposed by liquidity on an individual's savings behavior [10]. These papers, however, consider the economic nature of liquidity per se, leaving behind modeling of liquidity flows. On the contrary, numerous studies try to describe and measure liquidity via econometric models, including GARCH-based models [11] and ML algorithms [4]. Nevertheless, none of these studies makes use of unstructured data. At the same time, unstructured data has grown exponentially over the past decade [12], yet the number of studies that utilize unstructured data for liquidity forecasting is scarce.

Among them, however, text data has been used more successfully than others, with Natural Language Processing (NLP) being a fruitful instrument for handling it. Thus, the relationship between textual information and financial time series has been studied through sentiment dictionaries [13, 14] and machine learning methods [15, 16], including deep learning [17, 18]. Models based on news articles contribute to the prediction of stock trends [19], stock prices when combined with technical factors [20], and can be utilized to make profitable investment decisions [21]. Numerous studies successfully incorporate information from news into deep learning models to forecast stock volatility [22], the direction of stock markets [23], and stock trends [24]. Therefore, the usage of unstructured text data in trading volume forecasting looks promising. At the same time, quantifying unstructured textual data for further financial liquidity modeling procedures can take on various approaches. [25] demonstrate that daily Twitter activity and sentiment are associated with stock trading volume and can be used to predict next-day trading volume. [26] finds a positive correlation between the daily number of company mentions in the news and the daily transaction volume of a company's stock.

One fruitful idea, suggested by [27], is to use news diversity as a measure of uniformity of a news stream. The idea behind news diversity is quite simple. Suppose a significant event that can affect financial markets has occurred. In that case, the variety of topics covered in the articles among news publishers will be small since news publications will mostly cover topics connected with the corresponding event. In work [27], the LDA topic model [28] is used to extract probabilistic topics from the Financial Times news source, after which Shannon entropy as a natural measure of news diversity is applied to the distribution of topics. The diversity indicator is an additional exogenous feature in the endogenous ARMA(1, 1) model in predicting daily FTSE 100 trading volume changes. [27] find that diversity impacts trading volume. [29] examine the relationship between news diversity and financial market crashes and find that changes to news diversity are a valuable indicator of financial market crashes and recoveries. Overall, the research on news diversity is quite limited [30].

In our work, we expand the current concept of news diversity. First, we transform text data at three levels of abstraction: the level of words (occurrence and importance), the level of probabilistic topics (time-stable and time-varying), and the level of contextual embeddings. Secondly, we consider different types of entropy functions (parametric and non-parametric) to measure news diversity. Third, we examine the ability of various linear and non-linear ML models to extract predictively meaningful signals from entropy features. Fourth, we explore the scaling of news diversity ideas using a large number of time series of trading volumes in the stock market domain and a completely new financial liquidity domain for the applicability of news diversity - cash withdrawal volumes in the ATM network. In addition, we numerically compare an approach to text data preprocessing based on entropy functions with endogenous models, widely used shallow feature-based methods of text processing and lexicon-based sentiment models.

As mentioned earlier, this article also explores the impact of the news flow on the time series of cash withdrawals from ATMs. Currently, no studies are devoted to this topic, especially for emerging markets. Nevertheless, a more general task of forecasting cash demand in ATMs has been studied before. One of the first attempts to incorporate machine learning into ATM cash demand forecasting was the NN5 competition held in 2008. Since then, the dataset provided in the competition became a benchmark for comparison of forecasting approaches suggested by researchers [31–36]. However, the dataset is old and does not correspond to the dynamically developing banking industry. Moreover, it is unsuitable for examining news impact on cash demand since the amount of digitized English news corresponding to the dataset's time (1996-1998) is very scarce. On the other hand, [37–41] use more modern and non-public datasets to forecast daily cash demand for ATMs. However, all the mentioned studies have not used unstructured data, especially financial news, to forecast ATM cash demand. In this paper, we show that news features obtained using the proposed entropy approach provide a significant improvement in predicting changes in the volume of cash liquidity compared to models that do not use them.

3 Methodology

The methodology we propose for the task of predicting changes in financial liquidity flows is based on the concept of entropy from information theory and mathematical statistics. There, entropy is defined as a measure of uncertainty in the distribution of probabilities of a variable's outcome. Our variables of interest are calculated from business news flows treated at three levels of information generalization: at the level of words and at the level of documents, the latter being transformed into either probabilistic topic distributions or contextual vector representations (document embeddings). Once each variable is calculated, the value of entropy is computed over all variables aggregated by day or week, i.e. over all features related to words or texts that occurred in media sources in a given time unit. These entropy values thus constitute a feature space of a lower dimensionality and are then fed into predictive models. The stock market model predicts whether the volume change is higher than its median value, and the quality is measured as ROC-AUC of the respective curve. ATM withdrawal model predicts the difference of weekly withdrawal on a given date from its value from one year before, and the quality is measured as MAE of the respective regression. In the subsections below, we define the measures of uncertainty we use and the variables to which these measures are applied.

3.1 Entropy measurements

In our research, we test the efficiency of four entropy types: two classical entropies, Shannon's and Extropy as its dual function, and two parameterized entropies (Rényi and Tsallis) that have shown high efficiency in some other NLP-related tasks. For the two latter types, their respective hyper-parameters (α and q) are tuned through the time cross-validation procedure, resulting in the selection of the best model.

3.1.1 Shannon entropy

$$H(x) = -\sum_{n=1}^{N} P(x_n) \log P(x_n) \tag{1}$$

where $H(x)$ - Shannon entropy, x_n - event and $P(x_n)$ - probability of event.

3.1.2 Renyi entropy

$$H_\alpha(x) = \frac{1}{(1-\alpha)} \log \sum_{n=1}^{N} P^\alpha(x_n) \tag{2}$$

where $H_\alpha(x)$ - Rényi entropy, a generalization of the Shannon entropy, is a family of functionals used as a measure of the quantitative diversity, uncertainty, or randomness of some system, x_n - event and $P(x_n)$ - probability of an event, α - parameter.

3.1.3 Tsallis entropy

$$S_q(x) = \frac{k}{q-1}(1 - \sum_{n=1}^{N} P_i^q(x_n)) \tag{3}$$

where $S_q(x)$ - Tsallis entropy, a generalization of the standard Boltzmann–Gibbs entropy, x_n - event and $P(x_n)$ - probability of an event, where q is a parameter sometimes called the entropic index and k is a positive constant.

3.1.4 Extropy

$$E(x) = -\sum_{n=1}^{N} (1 - P(x_n)) \log(1 - P(x_n)) \tag{4}$$

where $E(x)$ - Extropy, the dual function for Shannon entropy [42], x_n - event and $P(x_n)$ - probability of event.

3.2 Entropy approach to word-level signals

In this subsection, we introduce two types of features calculated from the changes in either raw word frequency growth or weighted word frequencies in the entire news flow of a given time span.

3.2.1 Features based on raw word frequency growth

The idea behind using word frequency growth to calculate liquidity predictors is that usually it reflects the growth of media attention towards certain events happening for the first time or happening more frequently. As we investigate business news only, it is likely that the events covered in them may in turn reach financial actors and affect their liquidity-related behavior. Consider an n-units backward time window for each date, where n - parameter. Inside this time window, we count tokenized words that appear in the news in a given period. We ignore terms that appear in less than $\beta\%$ of the documents and more than $1 - \beta\%$ of the documents, where β - parameter. Next, we calculate percentage changes for the remaining terms and keep only the words with positive changes (i.e. only the words whose frequencies have grown). After finding terms with positive percentage changes for each time unit, we normalize these percentages by their time unit sum and calculate the entropy features (Shannon, Renyi, Tsallis, and Extropy).

Formally, for an arbitrary time unit i let

$$[d_{t_i-n}, d_{t_i}] \tag{5}$$

be the n-units backward time interval t_i, let

$$(w_1^i, w_2^i, \dots, w_{k_i}^i) \tag{6}$$

be the vocabulary of all words that occur in the given backward time interval for time unit i. The total number of words k_i may differ from one time unit to another. Let

$$(a_1^i, a_2^i, \dots, a_{k_i}^i) \tag{7}$$

be the corresponding number of instances of word $w_{k_i}^i$ in given backward interval for time unit i.

For each word w_j^i we calculate the change in the number of instances between backward interval for the time unit i and time unit $i - 1$, in percent:

$$r_j^i = \frac{a_j^i - a_j^{i-1}}{a_j^{i-1}}. \tag{8}$$

As mentioned earlier, we consider words with only positive r_j^i. Next, we normalize the change r as follows:

$$\rho_j^i = \frac{r_j^i}{\sum_i r_j^i} \tag{9}$$

and use its value to calculate entropies $H_i(\rho^i)$, $H_{\alpha,i}(\rho^i)$, $S_{q,i}(\rho^i)$ and $E_i(\rho^i)$.

3.2.2 Features based on smoothed word TF-IDF

As TF-IDF measures the relative importance of a word in a document, average word importance in a time span preceding a target liquidity-related event may indicate an event that has affected the target. To calculate such average importance, we compute words' TF-IDF with exponential smoothing, which is performed in several steps. First, we filter the most and the least frequent tokenized words, in the way described in the previous section. After that, we calculate TF and IDF for each word in each news article published in a given n-units backward time window. Then, we average TF values for each word across all articles in this window. Next, we calculate an exponentially weighted moving average within the same time span for both TF and IDF values of each word and normalize the resulting TF-IDF values by the sum for each time unit. As a final step, we calculate entropies using TF-IDF exponentially smoothed normalized values of each word as an input.

Formally, for each time unit and its n-units ahead time interval, we calculate TF parameter for every word t in each news article d:

$$\mathrm{TF}(t,d) = \frac{n_t}{\sum_i n_i} \tag{10}$$

where n_t is the count of a word t in the document d and $\sum_i n_i$ is the total word count in the document d. We then find the average value of TF values across all documents d published within n units time interval for each date:

$$\overline{\mathrm{TF}}(t,d) = \frac{\sum_{d \in D_i} \mathrm{TF}(t,d)}{|D_i|} \tag{11}$$

where D_i is a set of news articles released within n units time interval and $|D_i|$ is the count of those documents. IDF values are calculated as follows:

$$\mathrm{IDF}(t,D_i) = \log \frac{|D_i|}{|\{d_j \in D_i | t \in d_j\}|} \tag{12}$$

where $|\{d_j \in D_i | t \in d_j\}|$ is the count of documents (news articles) from the set D_i in which word t has appeared. Next, we find an exponentially moving average for $\overline{\mathrm{TF}}$ and IDF values. Let $[x_0, x_1, \dots, x_t]$ be an arbitrary time series value. Then the moving average is calculated as:

$$y_t = \frac{x_t + (1-\gamma)x_{t-1} + (1-\gamma)^2 x_{t-2} + \cdots + (1-\gamma)^t x_0}{1 + (1-\gamma) + (1-\gamma)^2 + \cdots + (1-\gamma)^t} \tag{13}$$

where γ is defined as $\gamma = 2/\mathrm{span} + 1$ and span is equal to n (time units). Afterwards we calculate the product of exponentially moving averages of $\overline{\mathrm{TF}}$ and IDF to find exponentially weighted TF-IDF for time unit i

$$\text{TF-IDF}_i^{smoothing} = \overline{\mathrm{TF}}_i^{smoothing} \times \mathrm{IDF}_i^{smoothing} \tag{14}$$

Those values are then normalized by the time unit sum and we get:

$$\overline{\text{TF-IDF}}_i = \frac{\text{TF-IDF}_i^{smoothing}}{\sum_j \text{TF-IDF}_j^{smoothing}} \tag{15}$$

This value is used to calculate entropies $H_i(\overline{\text{TF-IDF}}_i)$, $H_{\alpha,i}(\overline{\text{TF-IDF}})$, $S_{q,i}(\overline{\text{TF-IDF}}_i)$ and $E_i(\overline{\text{TF-IDF}}_i)$.

3.3 Features based on topical representations of documents

Event coverage happens at the level of texts, not words; this makes features aimed at the representation of documents in latent semantic spaces look promising for the task of liquidity prediction. To obtain a topical representation of a document, the preprocessed text is fed to the input of a probabilistic topic model; in this work, we use the classical LDA [28] and DTM [43] models.[1] Topic modeling procedure produces an n-dimensional vector of topics' probabilities for each document d in a given sample: $\theta^{(d)} = (\theta_{d,1}, \ldots, \theta_{d,n})$. The salience of topic T_j in each time unit t_i of the textual data stream D is statistically estimated as follows:

$$\Theta_i^j = \sum_{\forall d \text{ in } t_i} \theta_j^d \tag{16}$$

in the same manner, we normalize the topics saliencies to get probability distributions of semantic signals:

$$\overline{\Theta_i^j} = \frac{\Theta_i^j}{\sum_j \Theta_i^j} \tag{17}$$

Next, we calculate entropy features $H_i(\overline{\Theta_i})$, $H_{\alpha,i}(\overline{\Theta_i})$, $S_{q,i}(\overline{\Theta_i})$ and $E_i(\overline{\Theta_i})$.

3.4 Features based on contextual vector representations of documents

Finally, we introduce contextual document embeddings as text-level features. Embeddings aim to map words or other tokens from text corpus into vector space of predefined dimensionality [44]. To do this, we use Word2Vec, FastText, Doc2Vec, and BERT models on preprocessed texts. As a result, each document gets n-dimensional vector ϕ_i. After that, we apply Min-Max Scaling transformation [45] for embedding vectors:

$$\overline{\phi_i} = \frac{\phi_i - \phi_{min}}{\phi_{max} - \phi_{min}} \tag{18}$$

and calculate entropy features $H_i(\overline{\phi_i})$, $H_{\alpha,i}(\overline{\phi_i})$, $S_{q,i}(\overline{\phi_i})$ and $E_i(\overline{\phi_i})$.

3.5 Quality metrics

Here, we describe the mathematical formulation for the financial liquidity flow prediction problem. We also introduce metrics by which we evaluate the significance of our entropy-based features for assessing the impact of news on financial liquidity flows. We are more focused not on the absolute values of these metrics but on the relative changes in the metric, as compared to our baselines, both on average for the liquidity domain and in terms of the proportion of financial time series whose forecast has improved. Baseline models will be introduced later in the Experiments section.

3.5.1 Liquidity problem statement

Although all liquidity domains have many similarities, the specific behavior of their time series differs, which requires different formulations for the respective forecasting problems. The stock market is much more volatile than cash circulation; market behavior is

[1] We use LDA realization from gensim python package: https://radimrehurek.com/gensim/, C-language DTM realization: https://github.com/blei-lab/dtm.

influenced by many factors beyond the dynamics of the time series, the industrial calendar, and financial news. The task of predicting the exact value of trading volume is highly complicated. That is why we solve a two-class classification task to predict abnormal trading activity $\mathbb{1}_t^{abnormal}$ versus usual activity, instead of direct volume regression. Abnormal activity is understood as unusually deviant and is defined mathematically in the Stock market volume data and target variable subsection. The cash circulation domain is less volatile, but it experiences the impact of cyclic events, such as paydays and weekends, which often overshadows the effect of events covered in the news. Therefore, we choose to regress the weekly $YoY_{t,y}$ value which measures the difference of the current withdrawals volume from its value a year ago and is described in detail in ATM withdrawal data and target variable subsection. In both cases the chosen target variables allow us to focus on the prediction of changes in liquidity flows.

3.5.2 *Metrics for stock market volume*

We use the Receiving Operating Characteristics (ROC) and area under the curve (AUC) classification metrics to measure how well abnormal trading activity is detected. ROC-AUC is the universal metric for classification tasks as it estimates the model's predictive power for any given probability threshold.

3.5.3 *Metric for ATM withdrawals*

We evaluate the impact of news features through changes in the forecasting quality metric Mean Absolute Error in relation to the baseline model that does not use financial news data. MAE is described in the following formula:

$$MAE = \frac{\sum_{i=1}^{n} |y^i - y^i_{pred}|}{n} \tag{19}$$

where y^i is a target variable in i date, y^i_{pred} is a prediction for corresponding date i, and n is sample size.

Thus, the metrics we use are defined in the following way:

$$MAE_{improvement} = \frac{MAE_{baseline} - MAE_{exogen}}{MAE_{baseline}} * 100\% \tag{20}$$

where MAE_{exogen} is MAE value for a model with news features included in its feature space (extended baseline, sentiment baseline and entropy-approach in the Experiments section). $MAE_{endogen}$ is MAE value for a model with no additive financial news features (first baseline in the Experiments section).

4 Datasets and target variables

We collect two types of data (liquidity time series and corresponding economic news) for two application domains: stock market and ATM cash circulation.

4.1 Stock market volume data and target variable

We use time series data from the MOEX (Moscow Exchange) Russia Index for measuring liquidity in the stock market domain. MOEX index is a primary ruble-based capitalization-weighted index that tracks the performance of the largest and most liquid Russian companies from ten economy sectors. Thus, the trading volumes of shares included in this

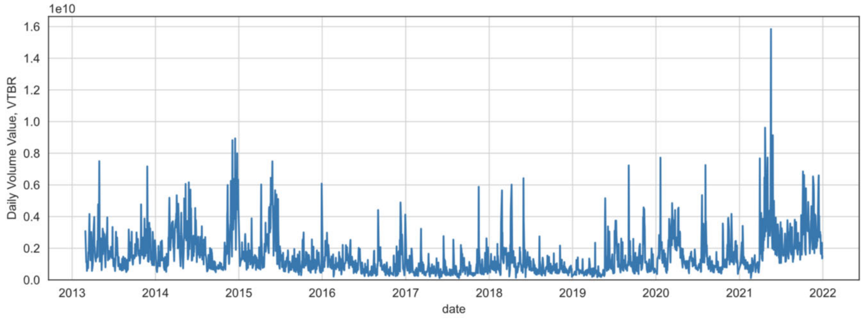


Figure 1 VTBR Volume

index reflect the distribution of the primary exchange liquidity in the Russian economy. Historical values of the index are available at https://www.moex.com/en/index/IMOEX. The constituents of the MOEX index considered in our research are listed below:

AFKS, AFLT, ALRS, CBOM, CHMF, FEES, GAZP, GMKN, HYDR, IRAO, LKOH, MAGN, MOEX, MTSS, NLMK, NVTK, PHOR, PIKK, PLZL, POLY, ROSN, RTKM, RUAL, SBER, SBERP, SNGS, SNGSP, TATN, TATNP, TRNFP, VTBR, YNDX

For example, Fig. 1 shows a graph of trading volumes in VTBR shares with typical bursts of liquidity, which are market reactions to certain economic events.

In this study, our main interest is in predicting abnormal trading activity. Abnormal trading activity may indicate both an idiosyncratic shock (the rise of trading activity due to rumors of mergers and acquisitions announcements) and a global event that can affect the market as a whole (e.g. COVID-19 pandemic). We define a particular day's trading volume as abnormal if its value exceeds the rolling median of trading volumes over the past n days. That is, day t has abnormal trading activity if

$$\mathbb{1}_t^{abnormal} = Volume_t - RM_{t-n} > 0 \tag{21}$$

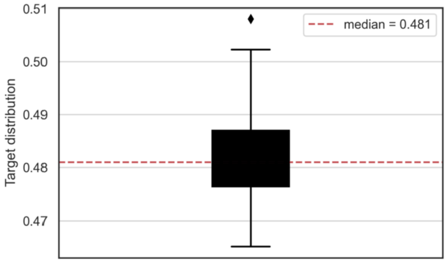
where Volume$_t$ is the trading volume at day t, RM_{t-n} is the trading volume rolling median over the past n days .[2] Further, as a time series in the stock market domain, we consider the indicator function $\mathbb{1}_t^{abnormal}$ of time. The distribution of $\mathbb{1}_t^{abnormal}$ across various trading volume time series for stocks included in the MOEX index is shown in Fig. 2. Notably, with this definition of abnormal trading activity, the classes of "abnormal" and "normal" trading activity are approximately balanced across different time series in our dataset.

4.2 ATM withdrawal data and target variable

In the cash circulation domain, we use the dataset from [37], which includes data on VTB bank ATMs in Moscow. VTB Bank is the second bank in terms of assets and the first bank in terms of authorized capital in Russia. The cash withdrawal values were anonymized for information security reasons, and 100 ATMs from January 2019 to December 2021 (which covers several waves of coronavirus spreading and lockdowns) were selected. The chosen ATMs have representative customer demand and were less affected by nonstationarities described in the article [37]. These ATMs also functioned during lockdowns (although the

[2] We set n equal to 30.

Figure 2 Abnormal trading activity distribution



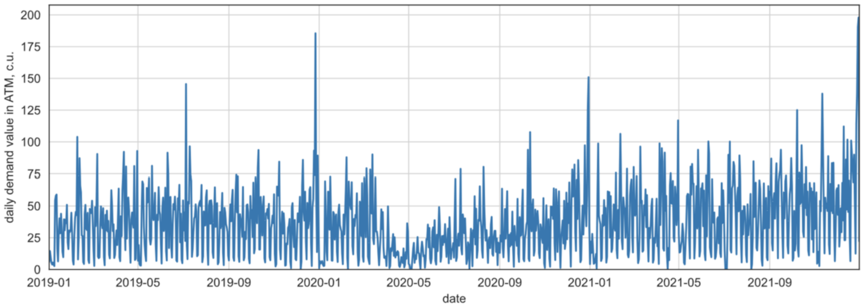


Figure 3 Example of ATM cash demand time-series

demand structure changed for some ATMs). The coronavirus lockdown in the spring of 2020 affected the demand for many ATMs most strongly (see Fig. 3).

Time series of cash withdrawals from ATMs are subject to the influence of weekly seasonality, public holidays, and regular events from the production calendar (e.g. paydays, advance days, and tax days). A detailed analysis of such influences is made in the article [37]. These cyclic events may obscure the effect of news and complicate the usage of daily or weekly withdrawal volume changes as target variables. To overcome these limitations, we apply the YoY (Year-Over-Year) transformation to each time series from the cash circulation domain. To do this, we determine the ordinal number of each week in a year, sum up all client withdrawals of cash liquidity within each week (thus removing within-week fluctuations), subtract the value of the liquidity sum of the week with the same ordinal number a year ago (we exclude monthly and annual seasonality) and give the expression as a share. The resulting transformation is described in the following formula [46]:

$$YoY_{t,y} = \frac{W_{t,y}}{W_{t,y-1}} - 1, \tag{22}$$

where $W_{t,y}$ is a sum of weekly withdrawals by ordinal number t and $W_{t,y-1}$ is a sum of withdrawals in the corresponding week with the same ordinal number in the previous year. Next, we consider the YoY function over time as a time series in the cash liquidity domain.

4.3 News data

As text data, we use news from four leading Russian-language media sources: Kommersant, RIA Novosti, Vedomosti, and Interfax. Kommersant (website: https://www.kommersant.ru) is a daily newspaper that is distributed throughout the country and fo-

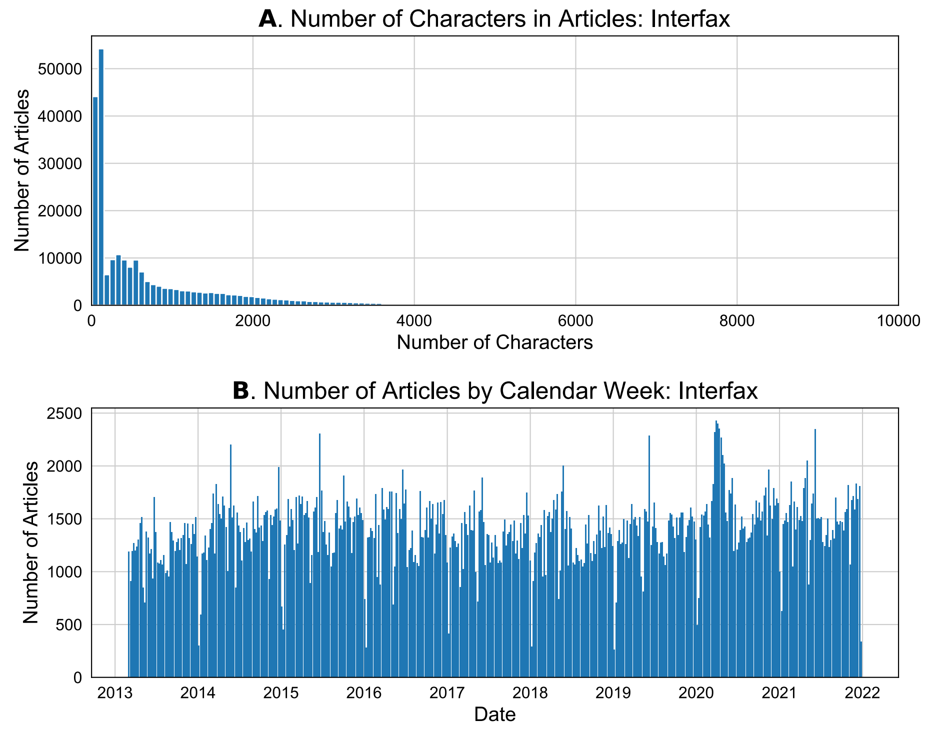


Figure 4 Interfax: The total number of articles by calendar week and the empirical distribution of the number of the characters

cuses on business and politics. Russia's leading state-owned news agency, RIA Novosti (Russian Information Agency; website: https://ria.ru), publishes news and commentary on social, political, economic, scientific, and financial topics. Vedomosti, a national daily newspaper focused on business, can be accessed at https://www.vedomosti.ru. Interfax (available on the website: https://interfax.com/) provides general and political news, business credit information, industry analysis, market data, and business solutions for risk, compliance, and credit management.

Of these sources, only Interfax news is used to predict liquidity in the stock market domain, as Interfax offers intensive intraday coverage of events related to MOEX Russia Index shares. It also covers every segment of the Russian economy, from resources and commodities to macroeconomic and corporate news. Our research considers only flash and express news as a more relevant and informative type of news for daily stock volume prediction. Also, shock events are more likely to be covered in short news. Figure 4 shows the main statistics about the Interfax newspaper.

Figure 4.A shows the empirical distribution of the number of characters in flash and express news from the Interfax agency. The mode of the presented distribution corresponds to shorter flash news, and the tail of the distribution corresponds to longer express news. Figure 4.B shows the distribution of the weekly number of articles over the considered period. The news number drops correspond to public holidays; the largest peak in 2020 associated with the COVID-19 pandemic.

We use the dataset from [16] with general and financial news from the rest three sources (from Kommersant, RIA Novosti, and Vedomosti) for liquidity prediction in the cash circulation domain. We use this dataset assuming that the decision to withdraw or deposit

Table 1 Summary Statistics of Collected News

News agency name	Number of articles	Dates
Kommersant	8653	01.01.2019 - 31.12.2021
RIA Novosti	44,827	01.01.2019 - 31.12.2021
Vedomosti	2232	01.01.2019 - 31.12.2021
Interfax	651,208	01.03.2013 - 31.12.2021

cash in an ATM network is more influenced by events that reflect the general economic news, and the impact of news on client withdrawals may occur with a delay of several days. The graphs with the empirical distribution of the number of characters and the number of articles by calendar week for Kommersant, RIA Novosti, and Vedomosti can be found in Appendix B. Table 1 illustrates the amount of collected data and the date intervals corresponding to it.

4.3.1 News data preprocessing pipeline

For constructing all types of features except BERT model embeddings, we used the same preprocessing that contained the following steps. Tokenization was performed with the natural language toolkit (NLTK package) in Python.[3] All tokens (words) were converted to the lowercase; punctuation, non-alphabet, and non-Cyrillic symbols, and non-Russian words were excluded.[4] Lemmatization, which converts words to their dictionary forms (lemmas), was performed on the remaining tokens with a Python wrapper for Yandex Mystem, a leading morphological analyzer for the Russian language.[5] The final step included removing tokens occurring in less than n_{below} documents and in more than n_{above} percent of the documents. We set n_{below} equal to 10 and n_{above} equal to 20%. For the BERT embeddings, we use the built-in tokenizer[6] to preprocess raw text.

5 Experiments

Our main experiments have been carried out via a grid search of the following structure (see Fig. 5). Bullets indicate lists of alternative feature sets or models, not used together. Since our focus is to test a new approach for feature engineering, we feed each of our feature sets into a set of standard machine learning models. As our tasks for stock market volume prediction and ATM withdrawal volume prediction were formulated differently, we apply different sets of models for each of them. The binary stock market volume (normal vs abnormal) is predicted with a set of classification models (logistic regression, GBM, SVM, and RF and 3-layered neural network used as classifiers), while YoY of ATM withdrawal volume is regressed on our features with linear regression, GBR, SVR, and RF and 3-layered neural network used in regression mode. Each of these ten models is fed with three types of features sets: first baseline, extended baseline, and the number of experimental sets based on our entropy approach. The first baseline includes only features from the time series and the industrial calendar (see First baseline: endogenous models subsection for more details). The extended baseline is represented by a group of feature sets, each consisting of the first baseline and textual features based on one of the popular contextual

[3] https://www.nltk.org.

[4] List of stop words is available from item 70 on https://www.nltk.org/nltk_data.

[5] https://pypi.org/project/pymystem3/.

[6] https://huggingface.co/cointegrated/rubert-tiny.

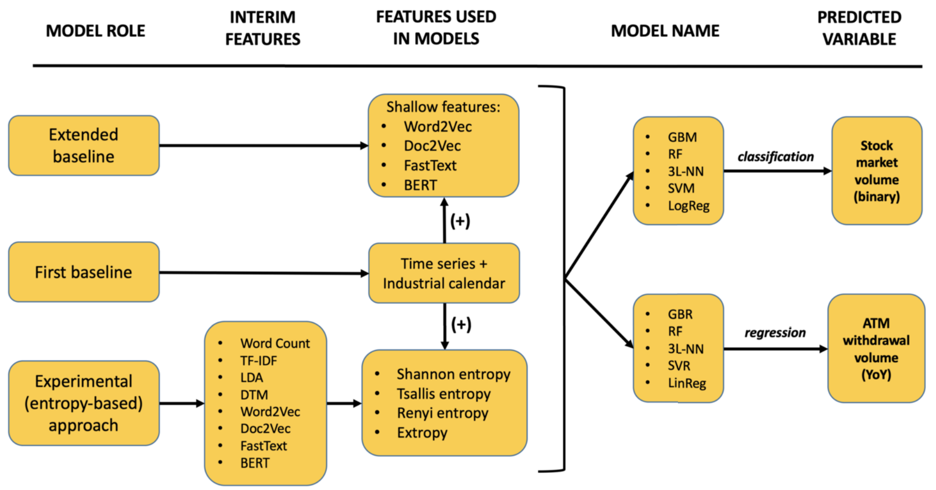


Figure 5 Structure of experiments

embeddings: Word2Vec, Doc2Vec, FastText, and BERT (find more details in Extended baseline: shallow feature enriched models subsection).

The process of feature construction for our experimental models was described in the Methodology section; in total, eight different entropy-based approaches to text preprocessing combined with four different types of entropy produce 32 feature sets. As each of them is also combined with the first baseline feature set, when each of them (plus the first baseline and four extended baseline feature sets) gets fed into five classification models and five regression models, we obtain 370 models that differ one from another either by feature constructing approach, or by the mathematical formalism of the model, or both. Of them, regression models are used to predict 100 time series in the ATM withdrawal domain, based on the text of each of the three news sources separately, and classification models are used to predict 32 time series in the stock market domain, based on Interfax news agency as a news source. In total, this produces 61,420 predictions.

In addition to the main experiments described above, we also conduct supplementary investigations to evaluate the applicability of models using news sentiment (see Sect. 5.4 for full description). Further, for regression tasks in the ATM cash withdrawal domain, we explore the applicability of the widely-used GARCH approach for modeling liquidity. This includes both its endogenous variant and its extension with the aforementioned methods for incorporating textual information (see Sect. 5.5 for further details).

5.1 First baseline: endogenous models

For both studied domains, we construct our first baseline using the same principle - endogenous economic features only, however, certain details of construction vary between the domains. In the stock market domain, as mentioned above, time series is highly volatile and does not have a clear seasonality or trend. Therefore, the task of isolating the influence of regularly repeating patterns of liquidity change is non-trivial. For these reasons, to build a baseline with acceptable quality, we use features from the industrial calendar encoded with one-hot transformation as factors. The resulting feature space which is then fed into one of the classification models includes the following features:

- Features based on the industrial calendar, encoded with one-hot transformation: day of the week, month, holiday, pre-holiday, after-holiday, last day of the month, pre-New Year, last workday of the month, weekend, holiday, the day before and after a holiday
- Rolling statistics: minimum, maximum, average, variance, median of the time series for 30 days before
- Rolling statistics grouped by weekday and month
- Lags: 5 previous values of the time series before the considered time interval

In the ATM withdrawals domain, we alter the time series by removing its regular data structure and the influence of the industrial calendar through weekly data aggregation and YoY transformation (described in the Datasets and target variables section). We use only five lags of corresponding time series for the baseline endogenous feature space.

5.2 Extended baseline: shallow feature enriched models

Our extended baseline combines the feature space for the first baseline described in the section above and news data transformed using one of the text embedding techniques as features. For news collection from each of the studied news sources (Kommersant, Vedomosti, RIA Novosti, and Interfax) we apply preprocessing techniques as described in the section Datasets and target variables. Further, each document is represented with one of the following approaches: Word2Vec, Doc2Vec, FastText, and BERT [7] (embeddings models were trained on the first two years of texts for each news outlet separately except the BERT model, which is a large language model pre-trained on a vast corpus of Russian-language texts). The resulting embeddings of news, where all the news for the same day (stock market) or week (ATM network) were treated as one document, were fed as the input features to the machine learning models described above.

5.3 Models with entropy-based features

Each of our experimental feature sets includes the feature space from the first baseline and a set of the features obtained by applying the entropy approach at the different levels of text data abstraction: words counts, probabilistic topics, and contextual embeddings.

At the word level, the proposed procedure has two main hyper-parameters: β and n. The first one is responsible for the influence of words on the final result through the frequency of their occurrence, and the second parameter is responsible for the locality of the considered economic context. Additionally, there is γ parameter that depends on n and is responsible for the decay in time of the change in the local context in the past to the current state of the economic system. All three hyper-parameters are selected independently based on the optimization of ROC-AUC or MAE metrics and the grid search on the training set for each liquidity time series.

At the topic level to determine the optimal number of topics for both LDA and DTM models, we build Roder's C_v metric for each media source separately [47]. Then, we find the number of topics at which the C_v curve reaches its maximum before flattening out (Figures B4-B7 in Appendix B). We set the frequency of changes in the probability of words within a topic in the DTM model to be equal to one month. We run both topic models on the training dataset and apply them to the test dataset.

[7] We use Word2Vec and Doc2vec realization from gensim python-package: https://radimrehurek.com/gensim/, FastText realization as python-package from https://fasttext.cc/, We use pre-trained embeddings of the BERT model for Russian language from Hugging Face website: https://huggingface.co/cointegrated/rubert-tiny.

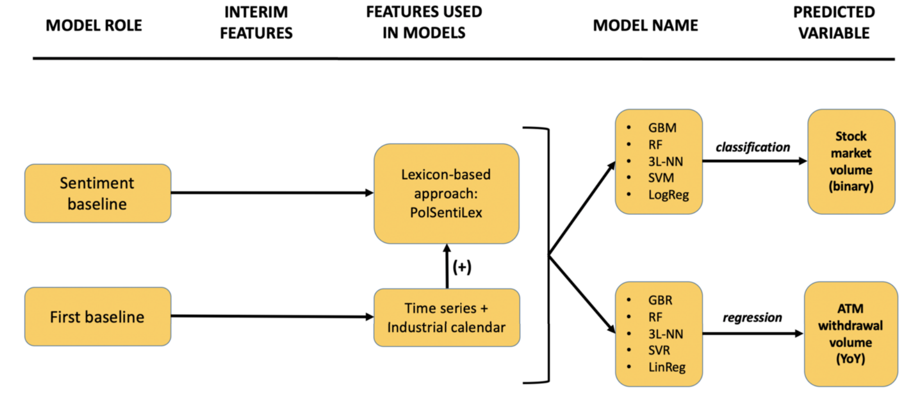


Figure 6 Structure of sentiment-based experiments

At the context level, we use the procedure for getting embeddings described in the Extended baseline: shallow feature enriched models subsection above.

Finally, the parameters α and q for the Renyi and Tsallis parametric entropies are selected separately on the training set for each news source but for all time series together (for saving computational resources).

5.4 Sentiment baseline: lexicon-based approach

As an additional baseline for comparing our proposed entropy-based features, we utilize PolSentiLex [48], one of the few existing sentiment lexicons for Russian language which, additionally, was specifically designed to analyze sentiment in social media content related to social and political issues. We expect that extensive usage of such lexicon in business news may increase their ability to affect financial behavior, e.g. by making individuals rush to trade their assets or to withdraw their cash.

We calculate sentiment at the level of individual news articles by averaging the sentiment scores of the words within each article. Subsequently, we aggregate these article-level sentiment scores through simple averaging: at the daily level for the stock market domain and at the weekly level for the ATM network domain. We then combine the feature space of the first baseline with the sentiment feature and use it as input for classification and regression models (see Fig. 6).

5.5 Econometric approach: GARCH model

In the ATM withdrawal volumes domain, where the task involves Year-over-Year (YoY) regression, we include the GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model as an additional predictive approach. GARCH models are widely recognized as a popular and effective instrument of modeling time series in finance and economics, particularly for capturing dynamics related to liquidity and volatility. Our choice is further motivated by the observation that our data often exhibit patterns of time-varying volatility and clustering, making them naturally suited to GARCH-based modeling. By applying the GARCH [8] model to this domain, we aim to provide an additional econometric perspective for addressing the prediction task.

[8] We use GARCH realization from arch python-package: https://github.com/bashtage/arch.

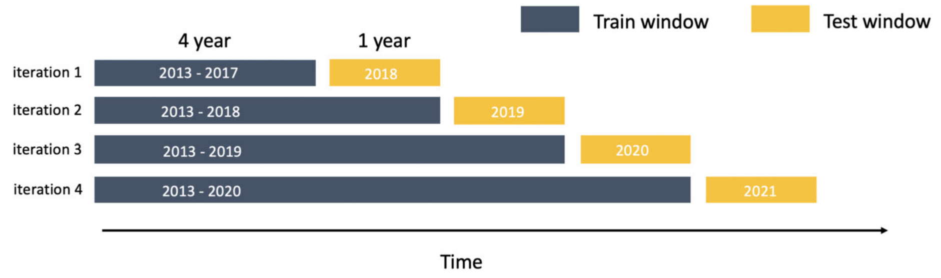


Figure 7 Train-test splitting: expanding window scheme

5.6 Evaluation procedure

An expanding window cross-validation scheme was used to account for the time structure of our data. With this technique, based on the time component, the dataset was split into n evenly distributed parts, each part having the size of the window, and the model was trained on the first k parts, with the subsequent part being the test set. The procedure continued by adding the test set to the training data (thus expanding training data by one part) and using the following subsequent sample as a new test set (see Fig. 7).

The window size was set equal to one year and the first four years were used as the initial train set with one year-ahead validation. Next, we trained machine learning models and applied them to test datasets, after which we stored all predicted values and used them to evaluate prediction quality metrics. Note that in the cash circulation domain, we have time series data only from 2019 to 2021; 2019 is used to calculate the YoY transformation; ML models are trained for 2020 and predict 2021.

To evaluate the statistical significance of the forecasting results, we employ the one-sided version of the Diebold-Mariano (DM) test [49][9] for regression tasks and Wilcoxon Signed-Ranks test [50] [10] for classification tasks. In this setup, endogenous models serve as the baseline, while the evaluated models are those incorporating textual data as features.

6 Discussion of numerical results

We analyze our numerical results separately for the stock market and for ATM cash withdrawal domains. In the analysis of each domain, we tackle the following issues: the impact of feature space extension with additional textual information on the results of forecasts, performance of the proposed entropy approach in comparison with the widely used shallow feature-based methods of text processing and lexicon-based sentiment, dependence of the results on the choice of entropy function (parametric / non-parametric), influence of the level of abstraction of text data representation and the model type (linear/non-linear models) on the prediction results, and, finally, impact of the text data source on the ability of ML approaches to extract significant signals for forecasting financial liquidity. Tables with all non-aggregated numerical results and their significance are presented in Appendix A.

[9] We use DM-test realization from epftoolbox python-package: https://github.com/jeslago/epftoolbox.

[10] We use Wilcoxon-test realization from scipy python-package: https://github.com/scipy/scipy.

Table 2 Top 10 approaches ranked by ROC-AUC and percent of time series with quality improvements in the stock market domain

Metric	Approach	Value
ROC-AUC	RF + Renyi + Word Counts	0.660
	RF + Tsallis + Word Counts	0.660
	RF + Shannon + Word Counts	0.660
	RF + Extropy + Word Counts	0.660
	RF + Renyi + FastText	0.659
	RF + Renyi + Word2Vec	0.659
	RF + Tsallis + FastText	0.659
	RF + Tsallis + Word2Vec	0.659
	RF + Shannon + Doc2Vec	0.659
	RF + Shannon + DTM	0.658
	Best first baseline	0.644
	Best extended baseline	0.644
	Best sentiment baseline	0.646
% of time series with quality improvements	3L-NN + Renyi + FastText	96.875
	3L-NN + Renyi + Doc2Vec	96.875
	3L-NN + Tsallis + Word Counts	96.875
	3L-NN + Shannon + Doc2Vec	96.875
	3L-NN + Extropy + LDA	96.875
	3L-NN + Extropy + Word Counts	96.875
	3L-NN + Tsallis + LDA	93.750
	3L-NN + Shannon + LDA	93.750
	3L-NN + Shannon + FastText	93.750
	3L-NN + Extropy + Word Counts	93.750
	Best extended baseline	68.750
	Best sentiment baseline	62.500

6.1 Stock market volumes

We begin our analysis by examining the best combinations of models and features in terms of the maximum quality for the ROC-AUC metric (our absolute metric) and the proportion of time series with forecast improvements compared to the endogenous baseline (our relative metric). These combinations are shown in Table 2. According to the ROC-AUC metric, the list of the best combinations includes only the Random Forest model. According to the second metric, the top combinations are represented only by the three-layer neural network model. The table also shows that incorporating textual data into the models enhances the prediction quality of abnormal trading activity. Moreover, the best results in terms of both metrics are achieved with entropy-based text features only.

Table 3 contains ten feature combinations with the highest mean values of absolute and relative quality metrics averaged across all models. It can be seen that all top feature sets presented in Table 3 are based on text information. Moreover, the majority of best-performing feature combinations are based on the entropy approach, in which vector text representation prevails. Indeed, six out of ten top-performing feature sets in terms of the first metric, and seven sets in terms of the second metric are constructed as entropies calculated from embeddings.

In contrast to the entropies obtained from embeddings, an extended baseline based on shallow feature-enriched models deteriorates the models' performance. At the same time, sentiment-based approaches provide slight improvements in forecasting accuracy but fall significantly short compared to entropy-based methods. Figure 8 shows forecasting results grouped by the feature construction approach.

Table 3 Top-10 features ranked by the mean value of ROC-AUC and percent of time series with quality improvements in the stock market domain

Metric	Feature name	Mean	Std
ROC-AUC	Renyi + Doc2Vec	0.6229	0.0362
	Renyi + FastText	0.6223	0.0361
	Tsallis + FastText	0.6223	0.0364
	Shannon + Word Count	0.6222	0.0385
	Tsallis + Word Count	0.6221	0.0375
	Renyi + BERT	0.6220	0.0367
	Extropy + DTM	0.6219	0.0362
	Extropy + Word Count	0.6219	0.0373
	Shannon + BERT	0.6218	0.0370
	Shannon + FastText	0.6218	0.0377
	Best extended baseline	0.6032	0.0456
	Best sentiment baseline	0.6022	0.0496
% of time series with quality improvements	Extropy + Word Count	80.62	17.59
	Shannon + FastText	79.38	17.48
	Renyi + Doc2Vec	79.38	17.76
	Extropy + DTM	78.75	13.87
	Shannon + Doc2Vec	78.75	16.15
	Shannon + BERT	78.75	18.80
	Shannon + Word Count	78.75	20.66
	Renyi + FastText	78.12	20.37
	Tsallis + Doc2Vec	77.50	15.92
	Tsallis + FastText	77.50	19.19
	Best extended baseline	25.00	34.30
	Best sentiment baseline	53.75	7.46

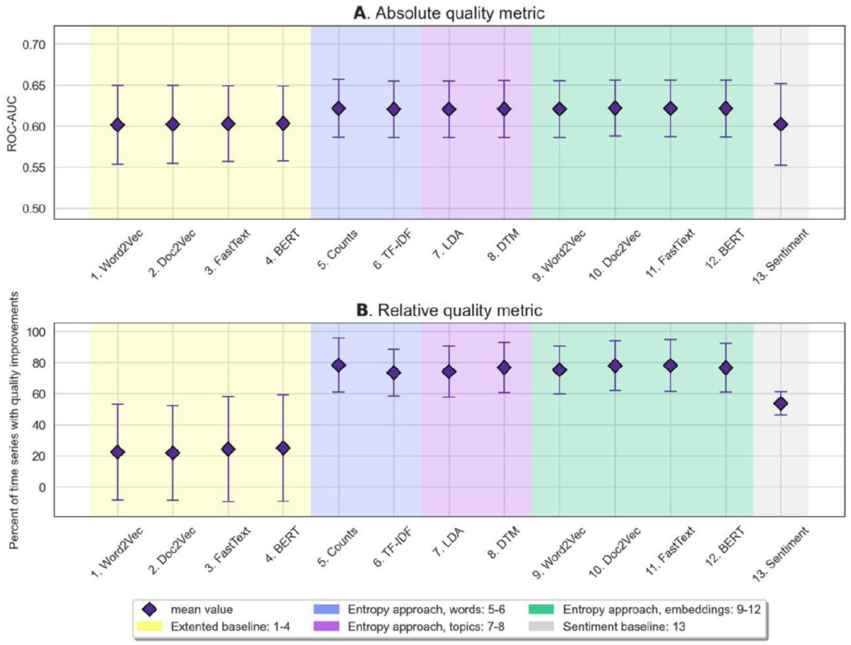


Figure 8 Forecasting results aggregated by feature construction approach, stock market domain

Figure 9 represents forecasting results aggregated by different entropy functions used. It demonstrates that neither individual entropy choice nor parametric vs non-parametric entropy choice has any effect on the forecasting results for the stock market domain by either absolute or relative metrics.

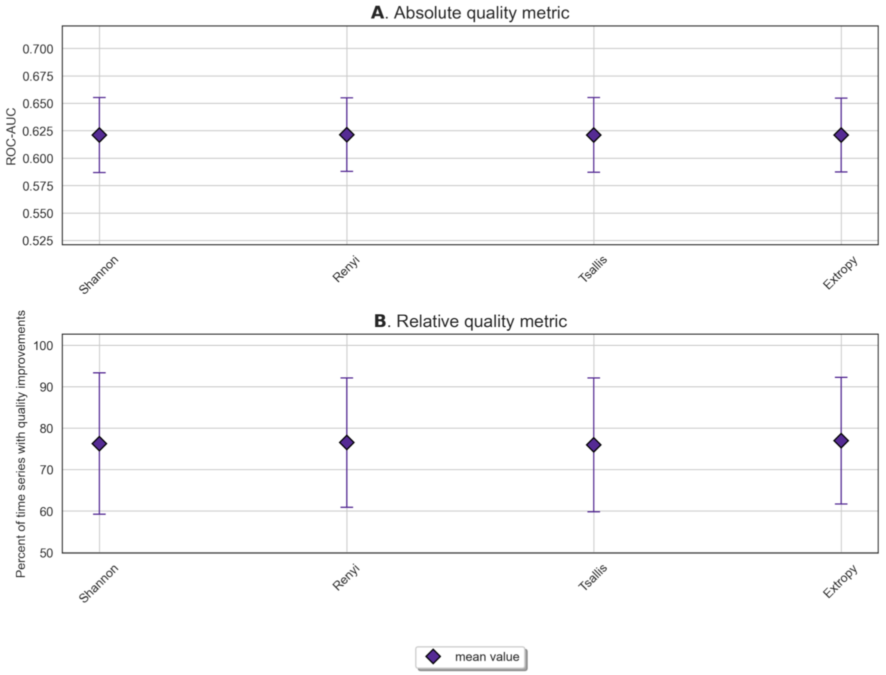


Figure 9 Forecasting results aggregated by entropy, stock market domain

Finally, we consider the impact of the model choice on the ability to predict abnormal trading volumes with entopy-based features. Figure 10 represents forecasting results aggregated by the chosen model in terms of absolute and relative quality metrics. According to the absolute metric (Fig. 10.A), the best results are achieved on models using decision trees (RF, average ROC-AUC score - 0.658). According to the relative metric (Fig. 10.B), the most significant gain is achieved on a three-layer neural network (on average 91% of time series demonstrate quality improvements), which, however, loses to LogReg in terms of the absolute metric (0.602 versus 0.623). At the same time, models based on decision trees seem to yield the best balance between the quality metrics, with relative quality reaching 84% of time series and absolute quality being the highest.

Overall, we can conclude that it is a combination of an entropy-based feature construction approach and a non-linear model that provides the highest prediction quality in the stock market domain, while the level of text data aggregation and the type of entropy are of little importance.

6.2 ATM withdrawal volumes

Passing to the ATM domain, we, too, start from the analysis of the best combinations of models and features. Table 4 presents these combinations sorted by the maximum MAE improvement and the proportion of time series with forecast improvements compared to the endogenous baseline. The results show the superiority of the text-based approaches over the endogenous model (first baseline), confirming that in this domain, too, news can increase models' predictive power. The top ten combinations in terms of MAE improvement are based exclusively on the three-layer neural networks models, while SVR, RF, and LinReg are among the ten models improving the largest proportion of time series. All best models use only entropy-based feature sets. In contrast to the stock market domain, the

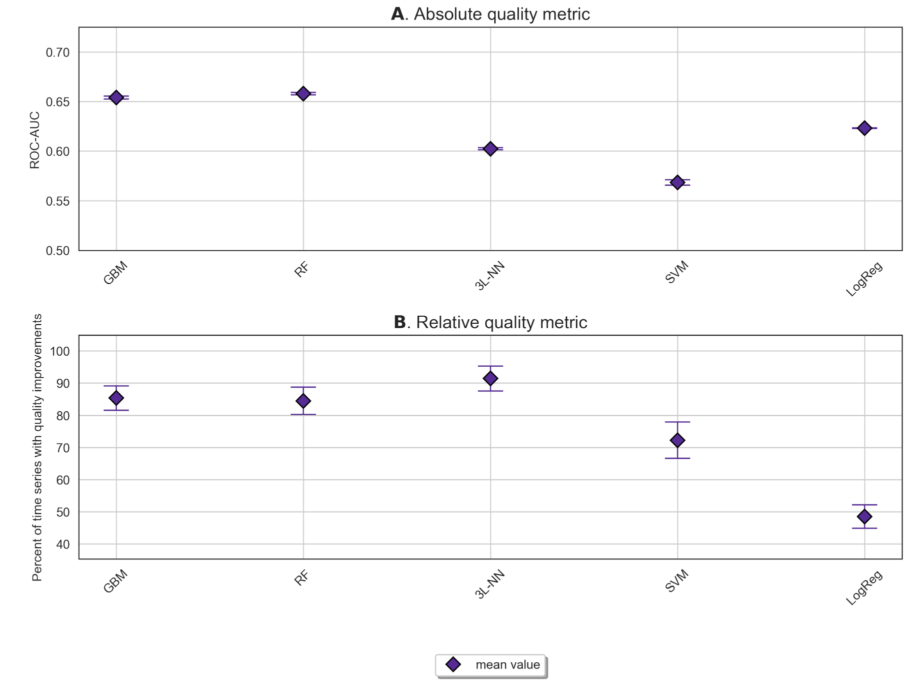


Figure 10 Forecasting results aggregated by model, stock market domain

best lexicon-based sentiment approach did make it into the top list for the relative metric, although it still significantly underperformed on the absolute metric, as compared to entropy-based approaches. Among news sources, RIA by far exceeds all the rest by its representation in both lists of the best models. The only exceptions are two Kommersant-based models appearing at the bottom of the list of models sorted by the proportion of time series improved.

In light of this visible inter-source variance, in our further analysis, the effects of individual news sources on model performance are given special attention. Table 5 contains the top five feature combinations for both metrics—relative and absolute—for each media source separately. We can see that, as in the stock market domain, only feature sets based on the entropy approach yield the best forecasting results, and no shallow feature-enriched approaches find their way into the top list. This is true for all news agencies. The situation is somewhat different when it comes to sentiment-based models: for Kommersant and RIA, the top results across both metrics are still dominated by entropy-based approaches, but for Vedomosti, sentiment-based models occupy the absolute top positions in terms of both metrics. Additionally, in contrast to the stock market domain, we observe multiple levels of abstraction in the source text data in this domain. It is also worth noting that the majority of the best-performing models are based on parametric entropies (ten out of 15 and nine out of 15 for MAE improvement and the proportion of time series improved, respectively).

Figure 11 shows forecasting results grouped by feature construction approach (shallow feature-based methods of text processing, lexicon-based sentiment approach and three types of entropy-based feature construction methods: word count, topic modeling, and embeddings). This representation is similar to Fig. 9, but visualizes each of the three news

Table 4 Top 10 approaches ranked by % of MAE gained and by percent of time series improved, ATM withdrawals domain

Metric	Approach	Value
MAE improvement, %	3L-NN + Tsallis + Word2Vec (Ria)	20.38
	3L-NN + Shannon + Word Counts (Ria)	20.07
	3L-NN + Tsallis + FastText (Ria)	19.35
	3L-NN + Renyi + FastText (Ria)	19.32
	3L-NN + Renyi + Doc2Vec (Ria)	19.26
	3L-NN + Tsallis + Word2Vec (Ria)	19.26
	3L-NN + Renyi + Word2Vec (Ria)	18.92
	3L-NN + Extropy + Word Counts (Ria)	18.69
	3L-NN + Renyi + Word Counts (Ria)	18.59
	3L-NN + Shannon + DTM (Ria)	18.42
	Best extended baseline	13.55
	Best sentiment baseline	13.03
% of time series with quality improvements	SVR + Renyi + FastText (Ria)	95.00
	SVR + Tsallis + FastText (Ria)	94.00
	SVR + Shannon + LDA (Ria)	94.00
	RF + Shannon + Word Counts (Ria)	93.00
	SVR + Renyi + LDA (Ria)	93.00
	SVR + Extropy + Word Counts (Ria)	93.00
	SVR + Extropy + LDA (Ria)	93.00
	RF + Shannon + Doc2Vec (Ria)	93.00
	LinReg + Tsallis + FastText (Kommersant)	93.00
	SVR + Tsallis + Word Counts (Kommersant)	93.00
	Best extended baseline	88.00
	Best sentiment baseline	93.00

sources separately. By total MAE improvement metric (Fig. 11.A), the extended baseline shows very noisy results with high variance, and we cannot conclude that on average it beats the endogenous model. Also, it can be seen that the results vary greatly between media - RIA Novosti shows better results than both Kommersant and Vedomosti for all entropy approaches (word, topic, and embedding levels). Globally, the top three best feature generation approaches are word counts, DTM, and Doc2Vec based on RIA Novosti. Further, as it can be seen from Fig. 11.B, the distribution of the proportions of improved time series by the levels of abstraction is similar to that in the stock exchange domain. The extended baseline models on average improve 51.55% of time series predicted with Kommersant data, 43.85% of those predicted with Vedomosti texts, and 54.0% of times series predicted with the news from RIA Novosti. The average improvement across all entropy-based approaches is 75.75%, 50.82% and 83.37% of time series, respectively, which is visibly higher. The results for the extended baseline in terms of the proportion of improved times series are as noisy as in terms of maximum MAE improvement. They are also far worse than the results for the entropy approaches. Globally, in the ATM domain, the best three approaches by the proportion of improved times series are exactly the same as by the maximum MAE improvement - word counts, DTM, and Doc2Vec combined with RIA Novosti. The effectiveness of the lexicon-based sentiment approach varies significantly across different sources. It performs poorest with RIA Novosti compared to other sources and methods, achieves performance comparable to other approaches for Kommersant, and delivers the best results for Vedomosti, surpassing both entropy-based and shallow feature-based methods. This variation is likely explained by dataset sizes, with Vedomosti being the smallest and RIA Novosti the largest collection. These findings align with previous observations that simpler approaches can outperform more complex ones when data

Table 5 Top 5 features ranked by the mean value of key metrics for three media sources, ATM withdrawals domain

Media name	Target variable	Feature name	Mean	Std
Kommersant	MAE improvement, %	Extropy + Doc2Vec	9.20	4.26
		Tsallis + LDA	8.46	4.27
		Shannon + Doc2Vec	8.45	4.06
		Tsallis + Words Count	8.25	4.31
		Renyi + LDA	8.24	3.35
		Best extended baseline	4.14	8.56
		Best sentiment baseline	6.33	3.96
	% of time series with quality improvements	Renyi + LDA	86.80	4.09
		Extropy + Doc2Vec	86.00	5.87
		Tsallis + Words Count	85.60	8.85
		Tsallis + LDA	84.80	6.42
		Renyi + Word2Vec	84.00	3.74
		Best extended baseline	64.60	24.54
		Best sentiment baseline	82	8.00
RIA	MAE improvement, %	Renyi + Words Count	14.15	3.37
		Tsallis + Words Count	14.01	2.00
		Tsallis + Word2Vec	13.73	4.14
		Renyi + Word2Vec	13.43	3.56
		Tsallis + FastText	0.26	14.96
		Best extended baseline	0.26	8.56
		Best sentiment baseline	–0.03	1.58
	% of time series with quality improvements	Renyi + FastText	88.40	4.93
		Tsallis + FastText	88.20	4.60
		Tsallis + TF-IDF	87.80	3.83
		Shannon + Words Count	87.60	5.55
		Extropy + Doc2Vec	87.40	5.41
		Best extended baseline	61.20	28.74
		Best sentiment baseline	47.80	9.58
Vedomosti	MAE improvement, %	Shannon + TF-IDF	3.07	5.55
		Tsallis + LDA	2.93	4.27
		Extropy + TF-IDF	2.77	5.30
		Renyi + LDA	2.61	3.75
		Shannon + Words Count	1.79	5.02
		Best extended baseline	–0.06	10.09
		Best sentiment baseline	3.29	3.66
	% of time series with quality improvements	Renyi + LDA	61.00	17.54
		Tsallis + LDA	59.60	18.89
		Shannon + TF-IDF	58.00	31.52
		Extropy + TF-IDF	57.40	29.47
		Shannon + Words Count	55.20	22.71
		Best extended baseline	51.20	26.14
		Best sentiment baseline	69.40	13.94

is scarce but struggle to compete when ample data is available. Overall, the sentiment approach does not yield the best results among the methods evaluated in this study.

Figure 12 shows how the values of forecasting metrics depend on the type of entropy and the news source. Our conclusions for the ATM domain are very different from those for the stock market domain where all entropies demonstrate similar results. Here, the two parametric entropies, Tsallis and Renyi, based on Kommersant and RIA Novosti data, obviously perform better than all other combinations in terms of both quality metrics. Although for Vedomosti the choice of entropy type does not significantly affect the forecasting result, on average, using parametric entropy further reduces MAE by 2 percentage points compared to methods that use non-parametric entropy. One possible explanation for this can be derived from the observation made by Lesche [51] who demonstrated that

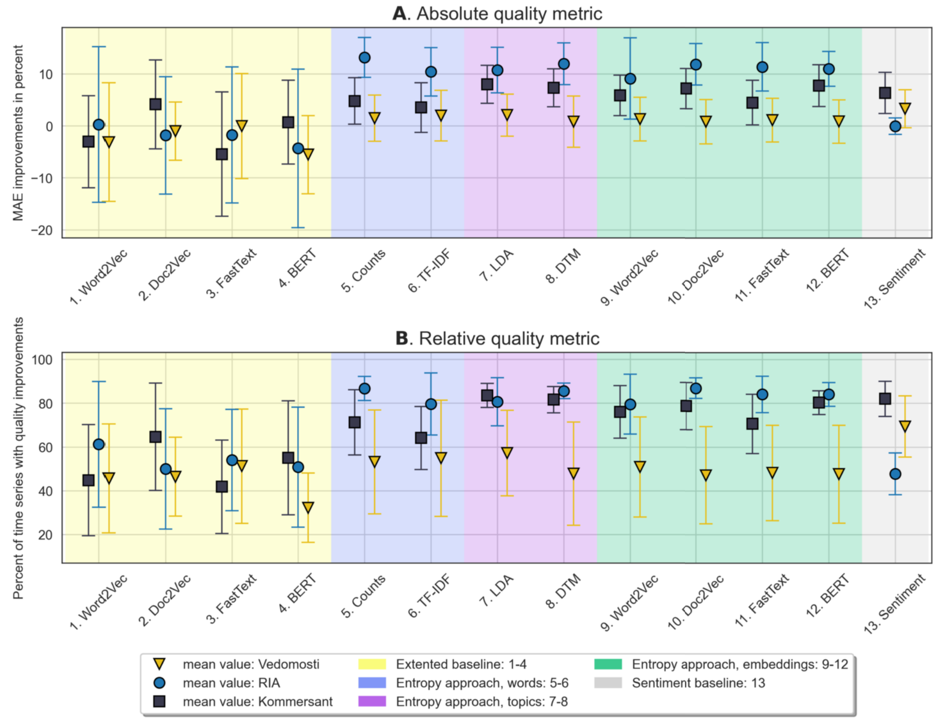


Figure 11 Forecasting results aggregated by level of abstraction, ATM network domain

Rényi entropy exhibits greater sensitivity to fluctuations in distributions compared to classical entropies, due to its parametric nature. This characteristic could potentially apply to other parametric entropies, offering them an advantage in capturing highly volatile distributions by adjusting the parameter q (referred to as the deformation parameter in statistical physics). In our study, stock price fluctuations appear to have experienced less dramatic change over time than those in cash withdrawals, which may explain why neither Rényi nor Tsallis entropies show a clear advantage over Shannon entropy in this prediction task. In contrast, the ATM domain is characterized by time series with significant structural changes over the studied period, influenced by external events such as the COVID-19 pandemic. As a result, parameterized entropies are able to better describe these data. Although this hypothesis is consistent with the observed performance differences, further research is necessary to substantiate the role of parametric entropy in addressing structural changes in time series forecasting.

The effect of the model choice on the values of both quality metrics can be seen in Fig. 13. According to the absolute metric (Fig. 13.A), the best model is 3L-NN based on RIA Novosti with a score of 15.95 %, while in terms of the relative metric (Fig. 13.B), the best performance is demonstrated by the combination of RF and RIA Novosti with the score of 89.66 %. Both combinations employ non-linear models, while the highest results of LinReg are 6.62% and 76.59 %, respectively. This suggests that non-linearity can contribute to the prediction quality improvement for methods that use the entropy approach.

All the above mentioned figures also speak in favor of a large variation between media sources in terms of their effect on the prediction quality. This is confirmed in Fig. 14. Entropy-based features constructed from RIA Novosti news on average allow for MAE

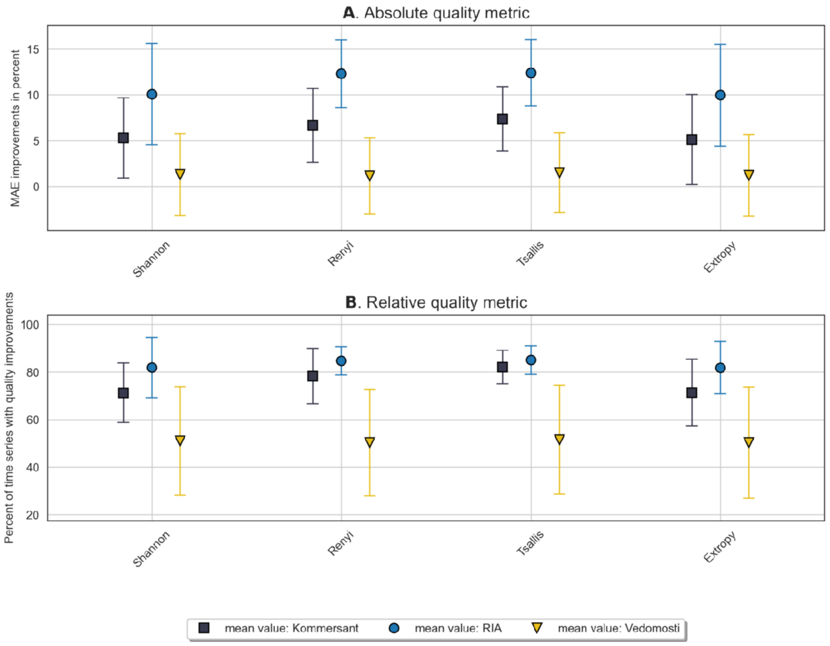


Figure 12 Forecasting results aggregated by entropy, ATM network domain

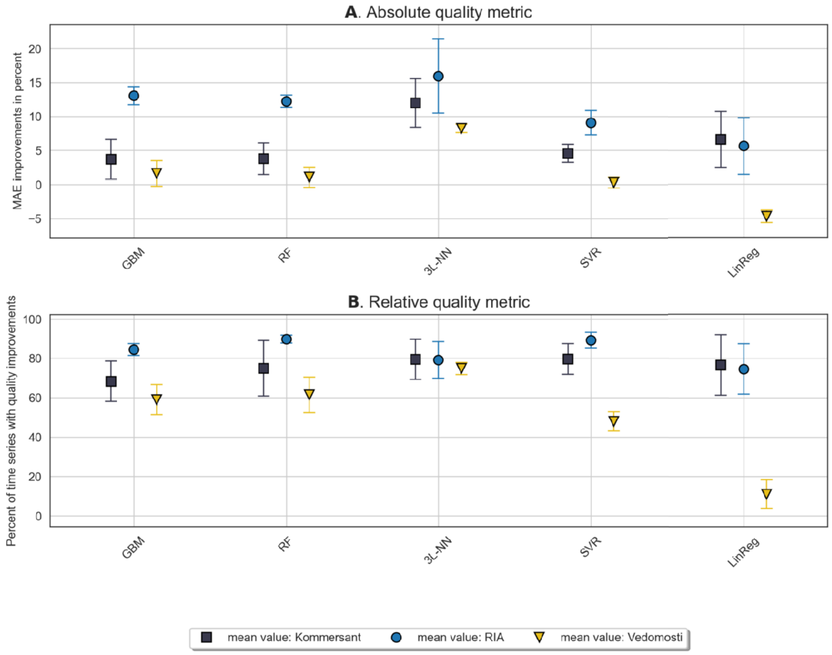


Figure 13 Forecasting results aggregated by model, ATM network domain

improvement by 11.2% thus outperforming models with features based on the other two sources. Specifically, Kommersant-based features yield MAE improvement by 6.11%, as compared to the first baseline, while using Vedomosti does not always improve the quality at all.

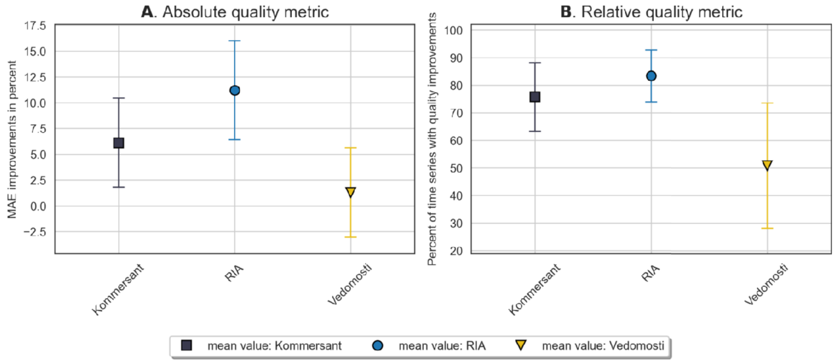


Figure 14 Forecasting results aggregated by media, ATM network domain

Table 6 Top 10 approaches ranked by % of MAE gained and % of time series improved, ATM withdrawals domain

Metric	Approach	Value
MAE improvement, %	GARCH + Renyi + TF-IDF (Kommersant)	25.55
	GARCH + Renyi + DTM (Ria)	24.48
	GARCH + Shannon + BERT (Kommersant)	23.38
	GARCH + Tsallis + FastText (Ria)	23.11
	GARCH + Renyi + BERT (Kommersant)	22.93
	GARCH + Tsallis + TF-IDF (Kommersant)	22.60
	GARCH + Extropy + BERT (Vedomosti)	22.46
	GARCH + Renyi + Counts (Ria)	22.26
	GARCH + Tsallis + LDA (Kommersant)	22.09
	GARCH + Renyi + LDA (Kommersant)	22.03
% of time series with quality improvements	GARCH + Extropy + DTM (Kommersant)	82.0
	GARCH + Tsallis + Doc2Vec (Kommersant)	81.0
	GARCH + Renyi + DTM (Ria)	80.0
	GARCH + Tsallis + DTM (Kommersant)	80.0
	GARCH + Renyi + DTM (Kommersant)	80.0
	GARCH + Tsallis + LDA (Kommersant)	79.0
	GARCH + Shannon + Doc2Vec (Kommersant)	79.0
	GARCH + Renyi + Doc2Vec (Kommersant)	78.0
	GARCH + Renyi + BERT (Kommersant)	77.0
	GARCH + Extropy + DTM (Ria)	77.0

6.2.1 GARCH model

Table 6 presents ten feature combinations with the highest values of absolute and relative quality metrics across all GARCH-based models. Notably, all entries in the table consist exclusively of models using the proposed entropy-based features. In terms of the MAE improvement metric, top 10 econometric approaches outperform the machine learning approaches listed in Table 4. However, when evaluating the percentage of time series with improved forecast quality, GARCH models fall behind ML models. This discrepancy suggests that GARCH-based models achieve significant improvements on certain time series but do not deliver consistent gains across the broader set thereof. Similarly, the results vary greatly for different feature sets (see Tables A.10 - A.11 in Appendix A). While some feature combinations yield outstanding improvements, others result in abnormal deterioration, and no pattern is traceable, and, moreover, the connection between feature type (entropy-based vs others) and the quality of the prediction is not traceable. All this indicates the low robustness of the econometric approach for the task under consideration.

Table 7 Performance of mixed top models combinations in the stock market domain

Metric	Approach	Value
ROC-AUC	(RF + Shannon + Word Counts) + (3L-NN + Renyi + FastText)	0.661
% of time series with quality improvements	(RF + Shannon + Word Counts) + (3L-NN + Renyi + FastText)	87.500

Table 8 Performance of mixed top feature combinations in the stock market domain

Metric	Feature name	Mean	Std
ROC-AUC	(Renyi + Doc2Vec) + (Extropy + Words count)	0.6023	0.0496
% of time series with quality improvements	(Renyi + Doc2Vec) + (Extropy + Words count)	53.76	7.47

6.3 Investigating feature and model synergies

In this section, we focus on the potential advantages of combining the top-performing features and ML models identified in our study, for the two liquidity domains under consideration. First, we evaluate the performance of blending the two best models based on absolute and relative metrics, respectively. This is done using a simple weighted average of the predictions, with the weights determined through cross-validation as outlined in the Evaluation procedure section. Second, we select the two feature types that exhibit the best average individual performance across all ML models under consideration, and then calculate mean performance values for the models using the two best feature types together.

Table 7 showcases the results of combining the two top-performing models in the stock market domain: the Random Forest model, trained using Shannon entropy based on word count (one of the four models exhibiting the highest ROC-AUC), and the 3L-NN model, trained with Rényi entropy utilizing contextual embeddings produced with FastText (the absolute leader according to the relative metric). In terms of this metric, blending the top models results in a decline in performance, whereas it slightly enhances performance when evaluated using ROC-AUC. Overall, the results remain fairly close to those achieved by models using one model and one feature type.

The best feature combination in the stock market domain for the absolute metric is the parametric Renyi entropy calculated using Doc2Vec embeddings, while for the relative metric, it is Extropy based on words count. The results of jointly leveraging these features across the ML models under consideration are presented in Table 8. As we can see, joint usage of these features lead to a performance decline according to both metrics as compared to the signle-type feature usage.

The results of blending the two best models for the ATM network domain are presented in Table 9. They include 3L-NN model trained with Tsallis entropy derived from RIA-Novosti-based Word2Vec, as the model that yielded the highest MAE improvement, and the SVR model trained with Renyi entropy derived from contextual embeddings produced by FastText (also based on RIA Novosti), as the best model in terms of the relative metric. In terms of both metrics blending the best models significantly degrades overall performance.

The best feature combination in the ATM domain for the absolute metric is the parametric Renyi entropy calculated using words count (based on RIA Novosti), while for the relative metric, it is Renyi entropy derived from contextual embeddings produced by FastText (based RIA Novosti). The results of jointly leveraging these features across the ML models under consideration are presented in Table 10. As we can see, joint usage of these features also worsens overall performance on both metrics.

Table 9 Performance of mixed top models combinations in the ATM network domain

Metric	Approach	Value
MAE improvement, %	(3L-NN + Tsallis + Word2Vec (Ria)) + (SVR + Renyi + FastText (RIA))	6.92
%of time series with quality improvements	(3L-NN + Tsallis + Word2Vec (Ria)) + (SVR + Renyi + FastText (RIA))	80.00

Table 10 Performance of mixed top feature combinations in the ATM network domain

Metric	Feature name	Mean	Std
MAE improvement, %	(Renyi + Words Count (Ria)) + (Renyi + FastText (Ria))	7.46	4.20
% of time series with quality improvements	(Renyi + Words Count (Ria)) + (Renyi + FastText (Ria))	77.80	13.66

Our findings demonstrate that, while the results vary across different domains, the combination of features and models either slightly improves performance on some metrics or significantly degrades it on others. This suggests that the best approach is to rely on individual features and models, as the current complexity of our framework is approaching a saturation point where additional complexity yields diminishing returns in predictive performance. It is important to note that our conclusions are based on the exploration of a local neighborhood of feature and model combinations, while a broader, more comprehensive investigation of combinations represents a promising direction for future studies.

7 Conclusion

In this paper, we proposed and explored an entropy-based text feature engineering approach for forecasting financial liquidity changes. The novelty of our work lies in several aspects. First, we introduce the use of economic news as unstructured data to predict changes in ATM cash withdrawal volumes, achieving visible improvement in forecasting quality. Second, we show that entropy-based approaches to generating textual features improve the quality of models for predicting changes in trading volumes on the stock exchange, as compared to models without textual data. Third, we, moreover, find that our entropy approach significantly outperforms the widely used shallow feature-based methods of text processing and lexicon-based sentiment approach for two different domains. Thus, our approach is promising as a potentially universal solution for text feature generation in many other liquidity domains. Fourth, while developing our entropy approach, we carry out extensive experiments and significantly broaden the spectrum of techniques allowing it to perform better, as compared to a few other approaches using entropies to predict financial liquidity flows [27, 29]. Specifically, the previous works used only LDA as the main model for generalizing information from raw text data. In contrast, we process text data at various levels of abstraction: at the level of changes in word frequency (relative frequency growth) and word importance (smoothed TF-IDF), at the level of time-stable (LDA) and time-varying (DTM) probabilistic topics, as well as at the level of context via embedding techniques (Word2Vec, Doc2Vec, FastText, and BERT). In addition, while the mentioned works considered only Shannon entropy, we use both non-parametric (Shannon, Extropy) and parametric (Renyi, Tsallis) functions. Finally, we explore the extent to which the type of machine learning model influences the ability to extract the most from entropy-based text features in terms of prediction quality, which has not been done before.

To prove the efficiency of the proposed feature engineering procedure, we have carried out experiments on the prediction of changes in the stock market trading volume of leading Russian companies and the volume of cash withdrawals in ATM network of a leading Russian bank. This was done with text data from the economic sections of leading Russian-language news media. We investigated several machine learning models: Gradient Boosting Machine (GBM), Random Forest (RF), 3-layer Neural Networks (3L-NN), Support Vector Machine (SVM), and Logistic Regression (LogReg) for classifying stock market volume (normal vs abnormal) and Gradient Boosting Regressor (GBR), Support Vector Regressor (SVR), Linear Regression (LinReg), Random Forest, and 3-layer Neural Networks in regression mode for predicting the difference between weekly withdrawal volume at each given date and the volume of withdrawals from the same ATM one year before. Additionally, we explored the applicability of the econometric approach based on Generalized Autoregressive Conditional Heteroskedasticity (GARCH) for predicting changes in ATM withdrawal volume.

We constructed and tested several types of feature spaces: endogenous features from the time series and the industrial calendar (first baseline), endogenous features enriched with text features obtained using Shallow feature-based methods of text processing (Word2Vec, Doc2Vec, FastText, BERT) (extended baseline), and endogenous features enriched with features obtained using the proposed entropy-based approach. Within the entropy-based approach, we studied four types of entropy functions (Shannon, Renyi, Tsallis, and Extropy), three different types of text abstraction (word level: words count, and TF-IDF technique; topic modeling: LDA and DTM topics, and embeddings: Word2Vec, Doc2Vec, FastText, and BERT). We also considered an extra baseline based on the sentiment lexicon from PolSentiLex [48], incorporating it into all the evaluated endogenous machine learning models as well as the econometric approach based on GARCH.

We tested 418 principles of feature space formation and model choice. Regression models were assessed for their ability to forecast 100 time series in the ATM withdrawal domain with text data from three news sources (Kommersant, Vedomosti, RIA Novosti) separately. Classification models were assessed for their ability to predict 32 time series in the stock market domain, based on Interfax news agency as a news source. In total, we generated 74,480 predictions. We assessed the quality of the constructed models using absolute and relative metrics. The absolute metric for the stock market volumes domain was the ROC-AUC score, for the ATM withdrawal volumes domain it was the proportion of change in the Mean Absolute Error in comparison to the baseline model that does not use financial news data. The proportion of financial time series whose forecast has improved was used as a relative metric for both considered domains.

In the task of predicting abnormal trading volumes on the Moscow Stock Exchange, the best result in absolute values according to the ROC-AUC metric is 0.660, which is achieved using the Random Forest model and the entropy approach to generating text features from the data on word frequency change. Two baseline models - the one using shallow feature-based methods of text processing (extended baseline) and the one using no text features at all (first baseline) - produce identical quality in terms of ROC-AUC (0.644). Thus, the extended baseline shows no improvement over the first baseline. The sentiment baseline, however, shows a slight but significant improvement, achieving an ROC-AUC of 0.646.

In terms of the proportion of time series with improved forecasting quality, the best result (96.875%) is achieved using a 3-layer Neural Network combined with entropy-based feature sets, which exceeds the best extended baseline result (68.750%) by 28.125 p.p.

Moreover, we find that the ten best feature spaces in terms of absolute and relative metrics consist only of those based on our proposed entropy approach. On average, shallow feature-based methods of text processing improve forecasting quality for 23% of time series, while the entropy approach shows an improvement in 76% of cases. Regarding the sentiment baseline, the best result for the relative metric is 62.5%, and on average, models utilizing sentiment improve forecasts for 54% of time series, which is also significantly lower than the results achieved by entropy-based approaches.

Additionally, in the stock market domain, as opposed to the ATM domain, we find no influence of the type of entropy (parametric vs non-parametric) on the forecasting results. Conversely, model type seems to affect prediction quality: non-linear models based on decision trees on average show higher quality in terms of both ROC-AUC score (0.658 from RF versus 0.623 from LogReg), and in terms of the fraction of time series with quality improvements. This points to the potential usefulness of non-linearity for quality improvement in the considered task.

In the problem of forecasting changes in ATM withdrawal volumes, the highest improvement in Mean Absolute Error (MAE) compared to the endogenous model is 25.55%, achieved using the GARCH approach with a feature space derived from the entropy method. For machine learning models, the best result is 20.38%, obtained with a 3-layer Neural Network, also utilizing a feature space based on the entropy approach. The best result for models using shallow feature-based methods of text processing is 13.55%. The best result for sentiment-based models is 13.03%.

In terms of the proportion of time series with improved forecasting quality, the best result is 95% achieved using a Support Vector Regressor based on entropy features, which exceeds the best extended baseline by 7 p.p. (95% versus 88%) and the best sentiment baseline by 2 p.p. (95% versus 93%). Moreover, the top five feature spaces in terms of both absolute and relative metrics consist only of the features generated with the proposed entropy approach for two of the three news sources considered. For the third news source, entropy-based features are also among the top performers but are outperformed by approaches based on lexicon-based sentiment which is likely related to the small size of Vedomosti dataset. In addition, GARCH-models with textual features and ML-models using shallow feature-based methods of text processing are characterized by a high degree of instability of results in terms of Mean Absolute Error improvements relative to the endogenous model.

The influence of the level of abstraction of source text data is uneven across news sources; globally the three best approaches for text transformation in the entropy approach are word counts, Doc2Vec, and DTM based on RIA Novosti. In contrast to the stock exchange domain, approaches using parametric entropies (Renyi, Tsallis) outperform those based on non-parametric entropies (Shannon, Extropy). Thus, in terms of MAE reduction compared to the endogenous model, parametric entropies provide an average improvement of 2 percentage points over their non-parametric counterparts. This advantage is likely due to the ability of parametric entropies to capture structural changes in distributions, such as those observed in our ATM data.

The choice of machine learning model also significantly influences forecasting results using the entropy-based text feature engineering approach. Like in the stock exchange domain, in the ATM domain we have suggestive evidence in favor of the positive role of model non-linearity for prediction quality. The best non-linear solution compares to that achieved by linear regression as 15.95% to 6.62% in terms of MAE improvement and as 89.66% to 76.59% in terms of the proportion of improved time series.

Our results also show a significant correlation between the choice of news source and the quality of the resulting models: the best results in forecasting are achieved by RIA Novosti and Kommersant news sources, but there is not always an improvement in the quality when Vedomosti is used. Summarizing the above, we can conclude that in the ATM withdrawals volume changes domain, non-linear models using the entropy-based text feature engineering approach with parametric entropy functions perform best.

Our approach has several practical implications. First, high-quality forecasts of abnormal trading volumes on the stock exchange allow optimizing distribution of the load on the exchange infrastructure, the developing of risk monitoring systems, and enriching trading strategies. Second, better prediction of ATM cash demand helps banks to optimize their ATM networks management, thereby improving user experience and minimizing operation costs. Third, a unified assessment of the impact of news on the most significant domains of financial liquidity allows for high-quality and interpretable stress testing in case of global macroeconomic shocks which also is a promising subject for further research. Finally, the proposed approach can extend to other markets and financial liquidity domains.

Appendix A

A.1 Sentiment baseline results

Table A.1 Results for sentiment baseline, stock market domain

Model	ROC-AUC	% of time series with quality improvements	Wilcoxon-test p-value
GBM	0.639	59	0.030
LogReg	0.626	62	0.060
3L-NN	0.568	53	0.000
RF	0.646	44	0.000
SVM	0.532	50	0.000

Table A.2 Results for sentiment baseline, ATM network domain

Media name	Model	MAE improvement, %	% of time series with quality improvements	DM-test p-value
Kommersant	GARCH	20.1761	64	0.046
	GBM	5.7822	77	0.000
	LinReg	5.8758	85	0.000
	3L-NN	13.0272	93	0.000
	RF	4.2435	83	0.000
	SVR	2.7310	72	0.000
RIA	GARCH	20.3186	46	0.043
	GBM	-2.3618	37	0.998
	LinReg	0.4243	58	0.076
	3L-NN	2.0272	57	0.024
	RF	-0.3022	40	0.805
	SVR	0.0443	47	0.463
Vedomosti	GARCH	21.3680	68	0.031
	GBM	1.8586	58	0.023
	LinReg	0.4587	57	0.082
	3L-NN	9.6513	91	0.000
	RF	1.6424	67	0.000
	SVR	2.8209	74	0.000

A.2 First and extended baselines results

Table A.3 Baseline results for stock market

Model	ROC-AUC
GBM	0.637
RF	0.644
3L-NN	0.568
SVM	0.538
LogReg	0.624

Table A.4 Results for raw embeddings, stock market domain

Metric	Model	Embedding			
ROC-AUC	GBM	0.6368	0.6368	0.6368	0.6368
	RF	0.6441	0.6441	0.6441	0.6441
	3L-NN	0.5699	0.5714	0.5724	0.5709
	SVM	0.5331	0.5338	0.5372	0.5398
	LogReg	0.6242	0.6242	0.6242	0.6242
% of time series with quality improvements	GBM	0	0	0	0
	RF	0	0	0	0
	3L-NN	59.38	62.50	68.75	65.64
	SVM	53.14	46.88	53.14	59.38
	LogReg	0	0	0	0

Table A.5 Results for raw embeddings, ATM network domain

Metric	Media name	Model	Embedding source			
			Word2Vec	Doc2Vec	FastText	Bert
MAE improvement, %	Kommersant	GBM	−7.42	5.81	−1.56	1.73
		RF	−5.85	1.82	−0.90	0.35
		3L-NN	7.74	13.55	−1.14	8.58
		SVR	4.16	8.65	2.94	5.39
		LinReg	−13.91	−9.14	−26.60	−12.52
	RIA	GBM	8.24	1.55	2.75	−2.51
		RF	4.46	−0.24	1.01	−3.09
		3L-NN	10.53	5.43	6.89	10.07
		SVR	4.13	5.60	5.36	3.68
		LinReg	−26.08	−21.53	−24.86	−29.85
	Vedomosti	GBM	−1.90	1.11	2.09	−6.34
		RF	−0.81	−0.23	0.03	−7.43
		3L-NN	6.69	5.11	10.82	0.69
		SVR	2.96	−0.91	3.35	2.09
		LinReg	−22.64	−10.18	−16.58	−16.74
%of time series with quality improvements	Kommersant	GBM	27	64	41	53
		RF	29	59	48	53
		3L-NN	72	85	45	78
		SVR	73	88	67	77
		LinReg	23	27	8	14
	RIA	GBM	75	54	60	49
		RF	70	51	60	41
		3L-NN	77	57	58	84
		SVR	74	82	77	68
		LinReg	10	6	15	12
	Vedomosti	GBM	40	51	49	30
		RF	49	52	49	21
		3L-NN	67	68	81	40
		SVR	66	42	66	55
		LinReg	6	19	11	15

A.3 Entropy-based text feature engineering approach results

Table A.6 Results for the stock market domain

Target variable	Entropy	Model	Topic level		Words level		Context level			
			LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
ROC-AUC	Shannon	GBM	0.653	0.654	**0.657***	0.654	0.652	0.653	0.655	0.655
		RF	0.656	**0.658****	**0.660*****	0.655	0.658	**0.659****	0.658	0.658
		3L-NN	**0.605***	0.602	0.602	0.602	0.602	0.603	**0.605***	0.603
		SVM	0.565	0.564	0.568	0.567	0.566	0.569	0.567	**0.570***
		LogReg	0.623	0.623	0.623	0.623	0.623	0.623	0.623	0.623
	Renyi	GBM	0.653	0.653	0.655	0.652	0.652	**0.657***	0.655	0.655
		RF	**0.657****	0.658	**0.660*****	**0.657****	**0.659****	0.658	**0.659****	0.658
		3L-NN	0.602	**0.603***	0.602	0.601	0.601	**0.603***	0.602	0.601
		SVM	0.567	0.569	0.565	0.567	0.572	**0.573***	**0.573***	0.572
		LogReg	0.623	0.623	0.623	**0.624***	0.623	0.623	0.623	**0.624***
	Tsallis	GBM	0.653	0.653	0.655	0.652	0.652	**0.657***	0.655	0.655
		RF	**0.657****	0.658	**0.660*****	**0.657****	**0.659****	0.658	**0.659****	0.658
		3L-NN	0.602	0.602	0.603	0.602	0.602	0.603	**0.604***	**0.604***
		SVM	0.568	0.564	0.570	0.568	0.568	0.568	**0.571***	0.566
		LogReg	0.623	0.623	0.623	0.623	0.623	0.623	0.622	0.623
	Extropy	GBM	0.653	**0.655***	0.654	**0.655***	0.652	0.654	0.652	**0.655***
		RF	0.656	**0.658****	**0.660*****	**0.657****	0.658	0.657	0.658	0.658
		3L-NN	**0.604***	0.602	0.602	0.602	0.602	0.600	0.602	0.601
		SVM	0.566	0.572	0.570	0.569	0.566	**0.573***	0.566	0.568
		LogReg	0.623	0.623	0.623	0.623	**0.624***	0.623	**0.624***	0.623
% of time series with quality improvements	Shannon	GBM	26	28	**30***	27	28	28	29	**30***
		RF	28	**29****	**29***	25	26	27	27	27
		3L-NN	30	28	30	28	27	**31*****	30	29
		SVM	19	22	22	19	23	22	**25***	**25***
		LogReg	14	14	15	15	15	**18***	16	15
	Renyi	GBM	25	28	**28***	27	26	27	**28***	27
		RF	25	**28***	**28***	25	27	**28***	27	26
		3L-NN	29	**29****	29	28	28	**31*****	**31*****	29
		SVM	23	22	23	22	25	25	25	**26***
		LogReg	16	16	15	**17***	15	16	14	15
	Tsallis	GBM	26	**28***	**28***	27	26	27	**28***	27
		RF	25	**28***	27	25	27	**28***	27	25
		3L-NN	30	29	**31*****	**29****	**30*****	30	29	28
		SVM	21	22	23	22	22	24	**26***	25
		LogReg	**17***	15	15	16	14	15	14	16
	Extropy	GBM	25	**28***	27	26	**28***	27	**28***	26
		RF	29	27	**30***	26	28	27	28	26
		3L-NN	**31*****	**29****	**31*****	**29****	26	29	28	**30****
		SVM	20	**24***	24	**24***	**24***	**24***	23	**24***
		LogReg	16	**18***	17	14	17	15	17	15

*: Best result in row.

**: Best result in column.

Table A.7 Kommersant, ATM network domain

Target variable	Entropy	Model	Topic level		Words level		Context level			
			LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
%MAE improvement, %	Shannon	GBM	5.44	4.00	–0.89	–0.05	1.05	6.24	–1.15	4.36
		RF	**5.86***	4.55	0.76	0.53	0.30	5.67	–0.36	3.75
		3L-NN	13.62	13.50	7.86	8.89	11.49	**14.77***	9.39	14.15
		SVR	4.28	4.10	2.90	2.03	4.72	5.28	3.68	**6.33***
		LinReg	9.34	9.08	4.91	–0.76	3.51	**10.29***	1.42	7.23
	Renyi	GBM	**6.68***	5.66	4.84	–0.72	6.41	6.24	4.36	5.65
		RF	5.41	5.46	5.07	2.19	5.38	5.00	4.82	**5.60*****
		3L-NN	**13.80***	**13.58****	12.16	9.23	11.86	11.48	**12.96****	**16.19****
		SVR	**6.48***	4.10	5.81	2.87	4.37	3.23	4.48	5.00
		LinReg	8.83	9.12	8.48	–3.47	8.84	0.54	**9.41***	8.71
	Tsallis	GBM	5.92	3.82	3.89	3.62	**6.39***	6.35	4.44	5.64
		RF	4.00	4.88	5.02	4.98	5.37	5.17	4.86	5.60
		3L-NN	**14.77****	13.44	**14.24****	**14.78****	**12.23****	6.02	10.08	**15.08***
		SVR	7.00	4.10	**7.09***	4.90	4.87	2.38	5.10	4.50
		LinReg	10.61	9.07	**11.00***	9.55	8.77	8.94	7.86	8.62
	Extropy	GBM	3.96	6.10	2.10	–0.71	0.61	**6.90***	–1.38	5.81
		RF	5.59	5.29	0.16	–0.88	0.33	**6.57***	–0.36	3.72
		3L-NN	14.37	13.43	–1.76	10.09	11.83	**16.18*****	8.58	15.08
		SVR	4.69	4.10	3.04	3.04	4.79	5.96	3.53	**6.77***
		LinReg	9.69	9.07	3.36	0.35	4.16	10.37	–2.25	7.09
% of time series with quality improvements	Shannon	GBM	**78***	69	49	53	56	73	56	69
		RF	86	79	61	50	56	**89***	50	73
		3L-NN	83	85	65	66	**86***	82	72	77
		SVR	79	82	73	57	78	84	72	**85***
		LinReg	83	87	64	53	62	**91***	62	76
	Renyi	GBM	**80***	71	72	49	79	74	71	78
		RF	**87***	83	81	59	**87*****	82	82	84
		3L-NN	**88***	85	80	63	**87****	79	85	**88*****
		SVR	**91*****	82	85	68	81	70	83	82
		LinReg	88	**88****	85	43	86	48	**89*****	87
	Tsallis	GBM	77	69	72	70	**78***	74	73	**78***
		RF	79	83	82	84	**87*****	83	82	84
		3L-NN	**90***	86	88	88	85	62	77	87
		SVR	**91****	82	**93*****	86	82	70	87	79
		LinReg	87	87	**93*****	**91****	86	83	81	87
	Extropy	GBM	71	75	48	56	58	77	55	**78****
		RF	85	84	54	48	57	**92*****	49	73
		3L-NN	85	**86***	46	73	**86***	84	69	79
		SVR	76	82	74	70	78	**87***	72	85
		LinReg	86	**87***	59	55	65	90	45	77

*: Best result in row.
**: Best result in column.

Table A.8 Vedomosti, ATM network domain

Target variable	Entropy	Model	Topic level		Words level		Context level			
			LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
MAE improvement, %	Shannon	GBM	0.57	0.51	3.29	**6.06***	1.22	−1.11	0.47	0.33
		RF	0.53	0.06	1.15	**6.75***	0.63	0.49	0.00	0.87
		3L-NN	7.96	**8.95***	**8.92****	7.87	8.16	**8.25****	**8.29****	7.68
		SVR	**1.12***	−1.02	0.54	−0.28	0.86	0.98	0.80	0.59
		LinReg	−4.72	−5.72	−4.97	−5.07	−4.81	−4.65	−4.80	−4.84
	Renyi	GBM	2.81	3.27	**3.60***	−2.82	2.12	1.31	0.95	0.84
		RF	**2.36***	1.15	1.16	0.16	0.45	1.00	0.66	0.18
		3L-NN	8.48	**9.00*****	7.22	8.03	7.75	6.83	7.86	7.45
		SVR	**1.26***	−1.04	−0.10	0.78	0.26	−1.24	0.79	0.00
		LinReg	−1.85	−5.66	−5.31	−4.17	−4.71	−5.32	−4.38	−5.07
	Tsallis	GBM	3.00	2.06	**3.64***	2.17	2.13	0.78	1.38	0.66
		RF	**1.98***	0.98	1.19	1.00	0.50	1.28	0.99	0.12
		3L-NN	**9.98*****	8.93	8.32	**9.62****	**8.20****	8.19	8.20	**8.20****
		SVR	**1.10***	−1.02	0.16	−0.85	0.74	0.74	0.74	0.74
		LinReg	−1.41	−5.78	−5.03	−4.16	−4.76	−4.76	−4.76	−4.76
	Extropy	GBM	2.43	0.12	2.74	**5.68***	2.00	−1.06	0.68	−0.39
		RF	1.31	−0.54	0.77	**5.40***	1.04	0.19	−0.23	0.65
		3L-NN	8.13	**8.93***	7.80	8.03	8.19	8.21	8.25	8.02
		SVR	**1.43***	−1.02	−0.57	−0.29	0.74	0.75	0.74	0.73
		LinReg	−5.21	−5.79	−5.22	−4.98	−4.77	−4.72	−4.78	−4.79
% of time series with quality improvements	Shannon	GBM	56	58	66	**73***	59	48	52	52
		RF	58	57	64	**90*****	60	58	51	56
		3L-NN	**79***	**75****	78	73	**74****	**75****	**76****	72
		SVR	**56***	40	49	44	50	51	49	49
		LinReg	22	8	19	10	8	9	9	8
	Renyi	GBM	61	61	**68***	46	64	56	51	58
		RF	**79***	69	63	61	58	52	57	54
		3L-NN	**77***	74	69	79	72	72	70	70
		SVR	50	40	45	**61***	47	41	49	46
		LinReg	38	8	8	9	10	6	12	4
	Tsallis	GBM	64	61	**69***	64	63	50	53	58
		RF	**71***	65	66	65	59	54	62	55
		3L-NN	**81*****	73	77	**81***	**74****	74	74	**74****
		SVR	49	40	**51***	42	50	50	50	50
		LinReg	33	8	6	11	9	9	9	8
	Extropy	GBM	61	51	65	**79***	63	49	57	51
		RF	66	47	64	**78***	63	54	50	57
		3L-NN	78	73	**79*****	74	**74****	**75****	74	72
		SVR	**55***	40	45	45	50	50	50	50
		LinReg	10	8	12	11	9	9	8	8

*: Best result in row.
**: Best result in column.

Table A.9 RIA, ATM network domain

Target variable	Entropy	Model	Topic level		Words level		Context level			
			LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
MAE improvement, %	Shannon	GBM	11.68	12.80	13.06	10.83	**13.28***	11.25	12.45	12.55
		RF	12.01	11.99	12.13	**12.88***	11.81	10.85	11.94	11.21
		3L-NN	**17.27****	**18.42****	**20.07*****	15.32	–2.95	18.16	15.27	**16.91****
		SVR	8.31	8.69	**10.98***	6.31	7.94	10.08	5.16	8.74
		LinReg	1.22	7.15	6.37	1.21	–6.09	**6.48***	3.47	6.34
	Renyi	GBM	11.72	13.16	16.52	14.17	13.83	13.28	**14.54***	12.52
		RF	12.40	12.20	13.11	13.17	**13.70***	11.46	13.42	11.49
		3L-NN	15.85	18.38	18.59	14.38	18.92	**19.26****	19.32	14.68
		SVR	10.22	8.68	**12.34***	7.76	9.53	10.29	11.06	8.97
		LinReg	9.56	7.16	10.18	4.11	**11.18***	7.65	7.38	**6.24***
	Tsallis	GBM	14.54	13.16	**15.77***	13.40	13.83	13.26	14.54	12.51
		RF	12.05	12.20	**13.79***	12.50	13.72	11.48	13.41	11.49
		3L-NN	12.93	18.39	16.32	**18.07****	**20.38*****	**19.26****	**19.35****	14.54
		SVR	8.92	8.69	**11.94***	9.88	9.49	10.35	11.14	8.99
		LinReg	7.19	7.12	**12.21***	7.67	11.24	7.43	7.37	6.19
	Extropy	GBM	12.26	13.25	**13.99***	10.91	13.39	11.17	11.58	12.65
		RF	12.06	**13.60***	11.96	12.90	11.64	**10.37***	11.70	11.28
		3L-NN	16.09	**18.41****	**18.69***	14.47	–4.23	18.02	15.02	16.90
		SVR	8.32	8.69	**11.01***	5.93	7.70	10.34	5.14	9.06
		LinReg	–0.04	**7.12***	4.67	2.10	–6.18	6.66	3.96	6.40
% of time series with quality improvements	Shannon	GBM	83	85	85	81	85	83	**88***	84
		RF	89	**91****	**93*****	86	89	**93*****	90	87
		3L-NN	79	81	87	80	53	88	75	**90*****
		SVR	**94*****	89	93	85	88	92	81	**90****
		LinReg	58	**83***	80	30	58	79	66	**83***
	Renyi	GBM	80	86	**90***	86	81	87	84	83
		RF	86	90	91	91	89	90	**92***	88
		3L-NN	70	82	86	84	79	84	**87***	71
		SVR	93	89	86	86	**91****	91	**95*****	84
		LinReg	78	83	81	70	80	81	**84***	81
	Tsallis	GBM	84	86	86	85	81	**87***	84	83
		RF	89	90	91	91	89	90	**92***	88
		3L-NN	67	81	75	83	81	83	**87***	71
		SVR	89	89	91	**92****	**91****	92	**94***	84
		LinReg	73	87	84	**88***	85	81	84	79
	Extropy	GBM	82	83	**90***	75	88	85	87	84
		RF	90	88	**93****	85	89	91	90	88
		3L-NN	79	81	86	77	51	**90***	74	89
		SVR	**93***	89	91	80	88	92	81	**90****
		LinReg	56	**83***	75	57	54	79	66	84

*: Best result in row.
**: Best result in column.

A.4 Econometric approach results

Table A.10 Results for raw embeddings and lexicon-based sentiment, ATM network domain

Metric	Media name	Model	Word2Vec	Doc2Vec	FastText	Bert	Sentiment
MAE improvement, %	Kommersant	GARCH	8.36	17.77	–9.20	5.99	20.18
	RIA		6.55	–134.04	17.20	8.52	20.32
	Vedomosti		3.33	–62.99	6.58	–1.02	21.37
% of time series with quality improvements	Kommersant	GARCH	20.0	40.0	9.0	8.0	64.0
	RIA		42.0	36.0	40.0	23.0	46.0
	Vedomosti		7.0	17.0	12.0	9.0	68.0

Table A.11 Results for GARCH model with entropy-based approach, ATM network domain

Media	Metric	Entropy	Topic level		Words level		Context level			
			LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
Kommersant	MAE improvement, %	Shannon	–12,712.384	16.001	–44.163	–11.235	–49.764	12.325	–15.412	23.382
		Renyi	22.026	20.621	–182.852	25.547	–31.323	–19.141	–108.504	22.934
		Tsallis	22.090	–71.191	12.832	22.595	6.604	4.376	8.238	19.795
		Extropy	11.322	15.181	–1.918	11.350	–31.054	–107.900	–58.384	20.426
	% of time series with quality improvements	Shannon	72	82	74	44	61	79	53	74
		Renyi	77	80	74	61	57	78	56	77
		Tsallis	79	80	66	62	64	81	51	74
		Extropy	74	82	76	60	71	71	43	64
Vedomosti	MAE improvement, %	Shannon	7.696	15.916	–1169.705	14.462	4.184	2.667	12.520	20.337
		Renyi	17.271	18.646	–434.900	17.534	3.147	–580.699	5.697	13.618
		Tsallis	15.993	14.877	–9,622,452 * 10^{12}	14.589	0.579	13.396	3.444	1.371
		Extropy	19.036	–303.916	1673.300	8.911	–7.147	12.017	10.114	22.458
	% of time series with quality improvements	Shannon	34	19	16	20	11	12	11	12
		Renyi	18	18	14	22	9	12	10	9
		Tsallis	13	18	13	19	9	13	10	11
		Extropy	27	19	23	19	9	9	10	10
RIA	MAE improvement, %	Shannon	–1.034	17.130	21.328	–38.644	–312.063	15.679	0.203	19.717
		Renyi	18.490	24.480	22.259	–32.509	0.517	15.562	11.553	13.911
		Tsallis	–18.304	17.262	18.475	–30.190	19.260	4.354	23.105	12.273
		Extropy	12.716	18.658	5.471	–50.390	17.913	14.090	–288.619	6.335
	% of time series with quality improvements	Shannon	59	77	44	40	46	68	57	54
		Renyi	62	80	47	47	39	68	64	66
		Tsallis	60	74	46	49	39	66	69	61
		Extropy	56	77	61	29	46	69	46	53

A.5 Statistical significance results

Table A.12 Wilcoxon-test p-values for raw embeddings, stock market domain

Model	Word2Vec	Doc2Vec	FastText	BERT
GBM	0.872	0.814	0.880	0.888
RF	0.067	0.066	0.064	0.067
3L-NN	0.000	0.000	0.000	0.000
SVM	0.970	0.970	0.816	0.000
LogReg	0.044	0.044	0.044	0.043

Table A.13 Wilcoxon-test p-values for entropy-based approach, stock market domain

Entropy	Model	Topic level		Words level		Context level			
		LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
Shannon	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LogReg	0.453	0.185	0.892	0.944	0.966	0.794	0.959	0.352
Renyi	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LogReg	0.953	0.185	0.697	0.315	0.713	0.183	0.332	0.41
Tsallis	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LogReg	0.448	0.185	0.624	0.484	0.397	0.415	0.374	0.562
Extropy	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LogReg	0.604	0.185	0.740	0.854	0.067	0.446	0.624	0.830

Table A.14 DM-test p-values for raw embeddings, ATM network domain

Media	Model	Word2Vec	Doc2Vec	FastText	BERT
Kommersant	GBM	1.0	0.0	0.839	0.147
	RF	1.0	0.025	0.8	0.371
	3L-NN	0.0	0.0	0.845	0.0
	SVR	0.0	0.0	0.0	0.0
	LinReg	1.0	1.0	1.0	1.0
RIA	GBM	0.0	0.169	0.039	0.916
	RF	0.0	0.604	0.138	1.0
	3L-NN	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0
	LinReg	1.0	1.0	1.0	1.0
Vedomosti	GBM	0.895	0.272	0.087	1.0
	RF	0.786	0.605	0.488	1.0
	3L-NN	0.0	0.0	0.0	0.266
	SVR	0.0	0.883	0.0	0.014
	LinReg	1.0	1.0	1.0	1.0

Table A.15 DM-test p-values for Kommersant, ATM network domain

Entropy	Model	Topic level		Words level		Context level			
		LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	Bert
Shannon	GBM	0.0	0.002	0.749	0.514	0.236	0.0	0.769	0.001
	RF	0.0	0.0	0.173	0.313	0.361	0.0	0.657	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.001	0.0	0.0	0.0	0.0
	LinReg	0.0	0.0	0.0	0.691	0.0	0.0	0.1	0.0
Renyi	GBM	0.0	0.0	0.0	0.682	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.003	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.0	0.0	0.0	0.991	0.0	0.29	0.0	0.0
Tsallis	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extropy	GBM	0.001	0.0	0.955	0.673	0.345	0.0	0.829	0.0
	RF	0.0	0.0	0.421	0.798	0.354	0.0	0.658	0.0
	3L-NN	0.0	0.0	0.869	0.0	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.0	0.0	0.0	0.403	0.0	0.0	0.966	0.0

Table A.16 DM-test p-values for Vedomosti, ATM network domain

Entropy	Model	Topic level		Words level		Context level			
		LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	Bert
Shannon	GBM	0.293	0.31	0.001	0.0	0.109	0.849	0.308	0.369
	RF	0.066	0.447	0.001	0.0	0.029	0.122	0.499	0.037
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.021	0.963	0.162	0.697	0.06	0.039	0.074	0.144
	LinReg	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Renyi	GBM	0.005	0.001	0.0	0.993	0.022	0.206	0.179	0.211
	RF	0.0	0.004	0.001	0.404	0.125	0.067	0.081	0.318
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.017	0.965	0.575	0.135	0.318	0.993	0.079	0.499
	LinReg	0.999	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Tsallis	GBM	0.002	0.01	0.0	0.009	0.022	0.311	0.083	0.263
	RF	0.0	0.001	0.001	0.001	0.097	0.03	0.013	0.377
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.028	0.963	0.391	0.959	0.09	0.09	0.09	0.09
	LinReg	0.997	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Extropy	GBM	0.006	0.453	0.003	0.0	0.026	0.817	0.245	0.644
	RF	0.0	0.852	0.02	0.0	0.001	0.338	0.703	0.112
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.006	0.963	0.849	0.699	0.088	0.087	0.089	0.093
	LinReg	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table A.17 DM-test p-values for RIA Novosti, ATM network domain

Entropy	Model	Topic level		Words level		Context level			
		LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	Bert
Shannon	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.834	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.198	0.0	0.0	1.0	0.989	0.0	0.002	0.0
Renyi	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.0	0.0	0.0	0.001	0.0	0.0	0.0	0.0
Tsallis	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extropy	GBM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	RF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3L-NN	0.0	0.0	0.0	0.0	0.908	0.0	0.0	0.0
	SVR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	LinReg	0.511	0.0	0.0	0.008	0.992	0.0	0.001	0.0

Table A.18 DM-test p-values for GARCH model with raw embeddings and lexicon-based sentiment, ATM network domain

Media name	Model	Word2Vec	Doc2Vec	FastText	BERT	Sentiment
Kommersant	GARCH	0.241	0.091	0.755	0.335	0.046
RIA		0.324	0.844	0.085	0.272	0.043
Vedomosti		0.404	0.860	0.308	0.534	0.031

Table A.19 DM-test p-values for GARCH model with entropy-based approach, ATM network domain

Media	Model	Entropy name	Topic level		Words level		Context level			
			LDA	DTM	Counts	TF-IDF	Word2Vec	Doc2Vec	FastText	BERT
Kommersant	GARCH	Shannon	0.860	0.072	0.913	0.865	0.883	0.225	0.767	0.029
		Renyi	0.043	0.046	0.848	0.041	0.836	0.721	0.978	0.018
		Tsallis	0.024	0.972	0.200	0.047	0.215	0.385	0.249	0.048
		Extropy	0.070	0.096	0.541	0.127	0.828	0.959	0.845	0.042
Vedomosti	GARCH	Shannon	0.171	0.096	0.930	0.103	0.393	0.329	0.144	0.136
		Renyi	0.075	0.066	0.826	0.084	0.359	0.839	0.334	0.139
		Tsallis	0.083	0.105	0.843	0.100	0.464	0.122	0.301	0.403
		Extropy	0.067	0.866	0.071	0.065	0.652	0.150	0.219	0.108
RIA	GARCH	Shannon	0.540	0.046	0.057	0.998	0.849	0.120	0.396	0.041
		Renyi	0.086	0.020	0.040	0.952	0.485	0.012	0.048	0.025
		Tsallis	0.745	0.030	0.056	0.823	0.073	0.396	0.033	0.196
		Extropy	0.170	0.081	0.259	0.990	0.075	0.123	0.917	0.078

Appendix B

B.1 News data figures

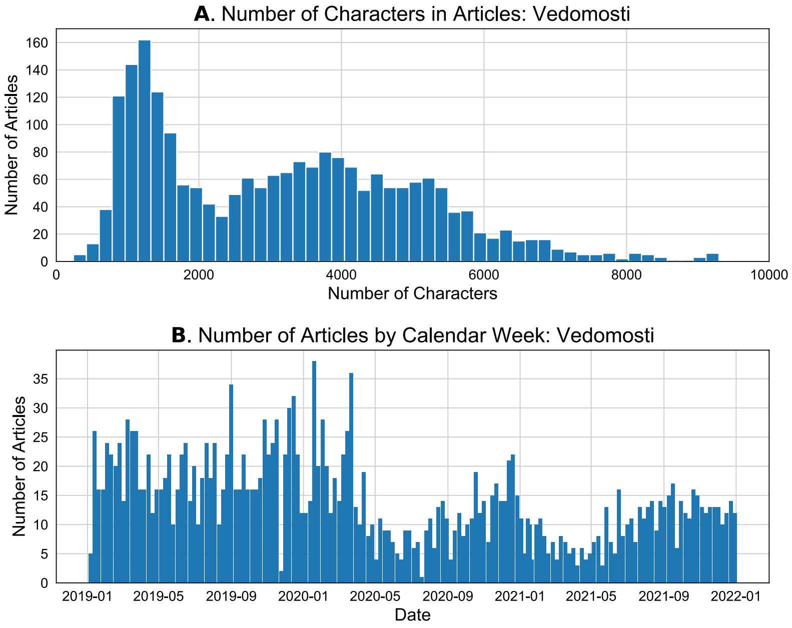


Figure B1 Vedomosti: The total number of articles by calendar week and the empirical distribution of the number of the symbols

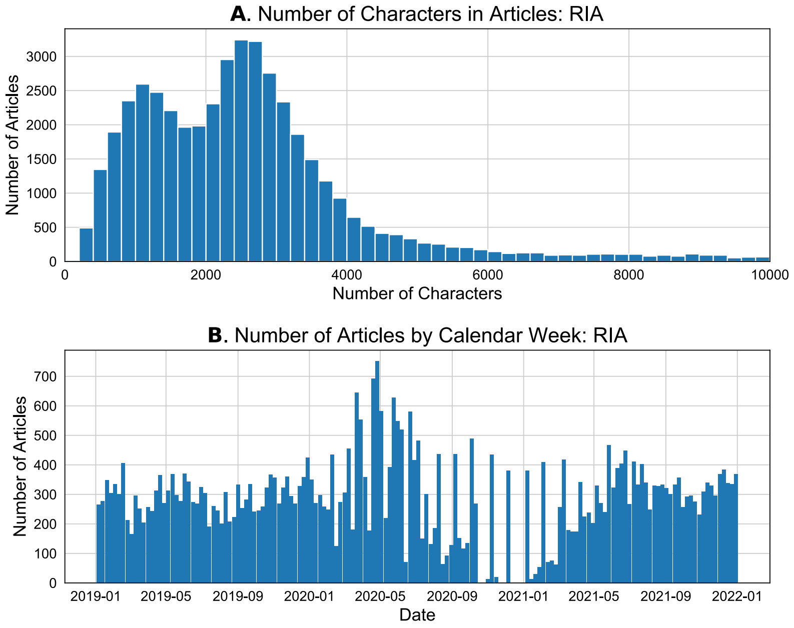


Figure B2 Ria: The total number of articles by calendar week and the empirical distribution of the number of the symbols

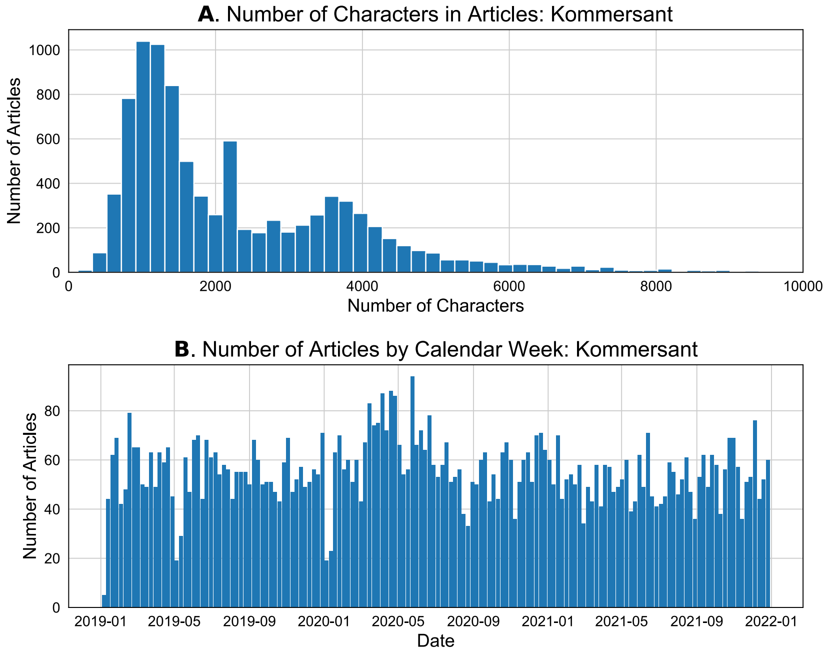


Figure B3 Kommersant: The total number of articles by calendar week and the empirical distribution of the number of the symbols

B.2 Topic modeling: CV score

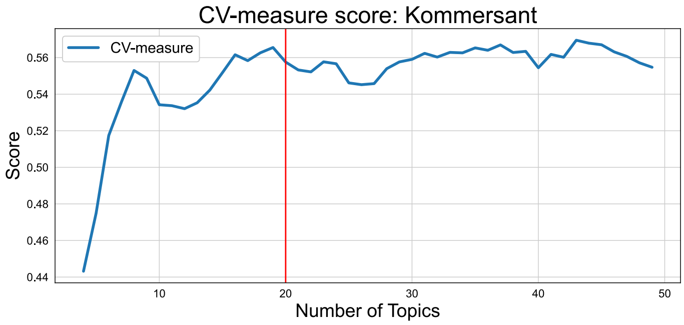


Figure B4 Kommersant: Dynamics of the CV-measure score

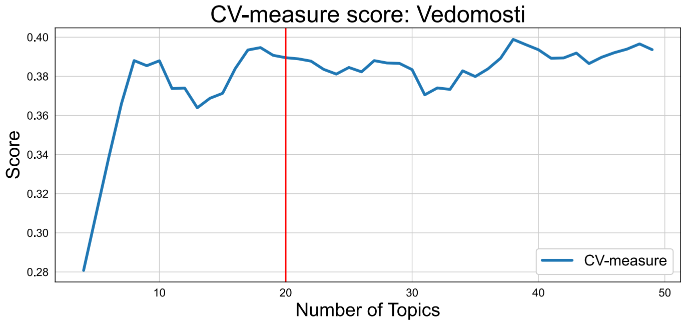


Figure B5 Vedomosti: Dynamics of the CV-measure score

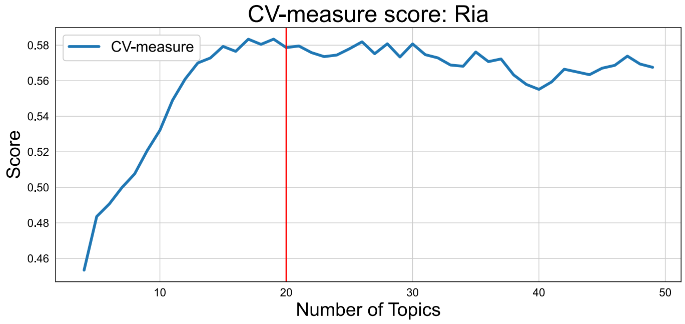


Figure B6 Ria: Dynamics of the CV-measure score

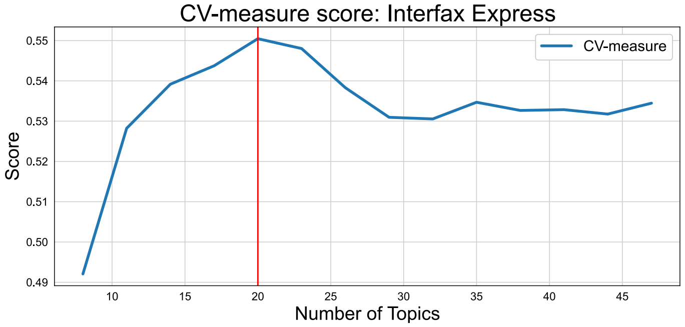


Figure B7 Interfax: Dynamics of the CV-measure score

Abbreviations

3L-NN, three-layered neural network; ARMA, Autoregressive Moving Average; ATM, automated teller machine; BERT, Bidirectional Encoder Representations from Transformers; COVID-19, Coronavirus disease of 2019; Doc2Vec, Documents to Vectors; DTM, dynamic topic model; FTSE, Financial Times Stock Exchange; GBM, Gradient boosting machine; GBR, Gradient boosting regressor; LDA, Latent Dirichlet Allocation; LinReg, linear regression; LogReg, logistic regression; MAE, mean absolute error; ML, machine learning; MOEX, Moscow Exchange; NLP, natural language processing; NLTK, Natural Language ToolKit; RF, Random forest; RIA Novosty, Russian Information Agency Novosty; ROC-AUC, Receiving Operating Characteristics (ROC) and area under the curve; SVM, Support vector machine; SVR, Support vector regressor; TF-IDF, term frequency times inverted document frequency; VTB, bank name (not an abbreviation); Word2Vec, Words to Vectors; YoY, year-over-year.

Acknowledgements
The authors are thankful to Sergei Koltcov for the advice on using certain types of entropies. The work of Aleksei Riabykh and Olessia Koltsova leading to these results has received funding from the Basic Research Program at the National Research University Higher School of Economics.

Author contributions
A. R.: research design, literature review, experiment design and performance, data analysis and interpretation, manuscript drafting and formatting, table and figure preparation, coding at all stages. I. S.: data retrieval, experiment performance, table and figure preparation, coding for these purposes, drafting manuscript sections. I. N.: experiment design and performance, table and figure preparation, coding for these purposes. D. S.: experiment design, manuscript editing. M. K.: experiment design, manuscript editing. O. K.: literature review, results interpretation, manuscript drafting and editing, table and figure preparation

Funding
The work of Aleksei Riabykh and Olessia Koltsova leading to these results has received funding from the Basic Research Program at the National Research University Higher School of Economics.

Data Availability
News texts are available on the respective media websites: https://www.kommersant.ru, https://ria.ru, https://www.vedomosti.ru and https://interfax.com/. MOEX index data is available at the Moscow Exchange website https://www.moex.com/en/index/IMOEX. ATM dataset is available from the first author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Ilias Suleimanov, Ilya Nagovitcyn, Denis Surzhko, and Maxim Konovalikhin are employees of VTB bank whose ATM dataset is are used in the research. VTB bank, however, has not funded this research and has no preference for certain research outcomes over others.

Author details
[1]Laboratory for Social and Cognitive Informatics, National Research University Higher School of Economics, 55/2 Sedova St., St. Petersburg, Russia. [2]Department of Data Analysis and Modeling, VTB Bank, Moscow, Russia.

Received: 12 June 2024 Accepted: 25 February 2025 Published online: 04 March 2025

References

1. De Meijer CR, Limburg L (2014) Intraday liquidity management and reporting: how to meet the challenges. J Risk Manag Financ Inst 7(4):395–408
2. Soprano A (2015) Liquidity management. In: A funding risk handbook, pp 1–191. https://doi.org/10.1002/9781119087946
3. Pflueger CE, Viceira LM (2016) Return predictability in the treasury market: real rates, inflation, and liquidity. In: Handbook of Fixed-Income Securities, pp 191–209
4. Guerra P, Castelli M, Nadine C-R (2022) Machine learning for liquidity risk modelling: a supervisory perspective. Econ Anal Policy 74:175–187
5. Climenta F, Momparlerb A, Carmona P (2019) Anticipating bank distress in the eurozone: an extreme gradient boosting approach. J Bus Res 101:885–896
6. Nopp C, Hanbury A (2015) Detecting risks in the banking system by sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 591–600
7. Leem B-H, Eum S-W (2021) Using text mining to measure mobile banking service quality. Ind Manag Data Syst 121(5):993–1007
8. Grossman S, Miller M (1988) Liquidity and market structure. J Finance 43(3):617–633
9. Myers SC, Rajan RG (1998) The paradox of liquidity. Q J Econ 113(3):733–771
10. Deaton A (1991) Savings and liquidity constraints. Econometrica 59(5):1221–1248
11. Zhao W, Gao Y, Wang M (2022) Measuring liquidity with return volatility: an analytical approach based on heavy-tailed censored-garch model. N Am J Econ Finance 62
12. Balducci B, Marinova D (2018) Unstructured data in marketing. J Acad Mark Sci 46(4):557–590
13. Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl-Based Syst 69:14–23
14. Picasso A, Merello S, Ma Y, Oneto L, Cambria E (2019) Technical analysis and sentiment embeddings for market trend prediction. Expert Syst Appl 135:60–70
15. Rundo F, Trenta F, Stallo AL, Battiato S (2019) Machine learning for quantitative finance applications: a survey. Appl Sci 9(24):5574
16. Riabykh A, Surzhko D, Konovalikhin M, Koltcov S (2022) Sttm: an efficient approach to estimating news impact on stock movement direction. PeerJ Comput Sci 8:1156

17. Matsubara T, Akita R, Uehara K (2018) Stock price prediction by deep neural generative model of news articles. IEICE Trans Inf Syst 101(4):901–908
18. Xu B, Zhang D, Zhang S, Li H, Lin H (2018) Stock market trend prediction using recurrent convolutional neural networks. In: CCF international conference on natural language processing and Chinese computing. Springer, Berlin, pp 166–177
19. Kalyani J, Bharathi P, Jyothi P, et al (2016) Stock trend prediction using news sentiment analysis. arXiv:1607.01958
20. Li X, Wu P, Wang W (2020) Incorporating stock prices and news sentiments for stock market prediction: a case of Hong Kong. Inf Process Manag 57(5):102212
21. Feuerriegel S, Prendinger H (2016) News-based trading strategies. Decis Support Syst 90:65–74
22. Wang Y, Liu H, Guo Q, Xie S, Zhang X (2019) Stock volatility prediction by hybrid neural network. IEEE Access 7:154524–154534
23. Hu H, Tang L, Zhang S, Wang H (2018) Predicting the direction of stock markets using optimized neural networks with Google trends. Neurocomputing 285:188–195
24. Hu Z, Liu W, Bian J, Liu X, Liu T-Y (2018) Listening to chaotic whispers: a deep learning framework for news-oriented stock trend prediction. In: Proceedings of the eleventh ACM international conference on web search and data mining, pp 261–269
25. Tan SD, Tas O (2021) Social media sentiment in international stock returns and trading activity. J Behav Finance 22(2):221–234. https://doi.org/10.1080/15427560.2020.1772261
26. Alanyali M, Moat HS, Preis T (2013) Quantifying the relationship between financial news and the stock market. Sci Rep 3(1):1–6
27. Curme C, Zhuo Y, Moat HS, Preis T (2017) Quantifying the diversity of news around stock market moves. Chester Curme, Ying Daisy Zhuo, Helen Susannah Moat and Tobias Preis, Quantifying the diversity of news around stock market moves. J Netw Theory Finance 3(1):1–20
28. Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
29. Boubaker S, Liu Z, Zhai L (2021) Big data, news diversity and financial market crash. Technol Forecast Soc Change 168:120755. https://doi.org/10.1016/j.techfore.2021.120755
30. Hendrickx J, Ballon P, Ranaivoson H (2020) Dissecting news diversity: an integrated conceptual framework. Journalism 23(8):1751–1769
31. Andrawis RR, Atiya AF, El-Shishiny H (2011) Forecast combinations of computational intelligence and linear models for the nn5 time series forecasting competition. Int J Forecast 27(3):672–688
32. Wichard JD (2011) Forecasting the nn5 time series with hybrid models. Int J Forecast 27(3):700–707
33. Taieb SB, Bontempi G, Atiya AF, Sorjamaa A (2012) A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. Expert Syst Appl 39(8):7067–7083
34. Kamini V, Ravi V, Kumar DN (2014) Chaotic time series analysis with neural networks to forecast cash demand in atms. In: 2014 IEEE international conference on computational intelligence and computing research. IEEE Press, New York, pp 1–5
35. Venkatesh K, Ravi V, Prinzie A, Poel D (2014) Cash demand forecasting in atms by clustering and neural networks. Eur J Oper Res 232(2):383–392
36. Jadwal PK, Jain S, Gupta U, Khanna P (2017) K-means clustering with neural networks for atm cash repository prediction. In: International conference on information and communication technology for intelligent systems. Springer, Berlin, pp 588–596
37. Riabykh A, Suleimanov I, Surzhko D, Konovalikhin M, Ryazanov V (2022) Atm cash flow prediction using local and global model approaches in cash management optimization. Pattern Recognit Image Anal 32(4):803–820. https://doi.org/10.1134/S1054661822040113
38. Serengil S, Ozpinar A (2019) Atm cash flow prediction and replenishment optimization with ANN. Int J Eng Res Dev 11:402–408. https://doi.org/10.29137/umagd.484670
39. Arabani SP, Komleh HE (2019) The improvement of forecasting atms cash demand of Iran banking network using convolutional neural network. Arab J Sci Eng 44(4):3733–3743
40. Rafi M, Wahab MT, Khan MB, Raza H (2020) Atm cash prediction using time series approach. In: 2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET). IEEE Press, New York, pp 1–6
41. Lázaro JL, Jiménez ÁB, Takeda A (2018) Improving cash logistics in bank branches by coupling machine learning and robust optimization. Expert Syst Appl 92:236–255
42. Lad F, Sanfilippo G, Agrò G (2015) Extropy: complementary dual of entropy. Stat Sci 30(1):40–58. https://doi.org/10.1214/14-STS430
43. Blei D, Lafferty J (2006) Dynamic topic models. In: ICML 2006 - proceedings of the 23rd international conference on machine learning, pp 113–120
44. Bengio Y, Ducharme R, Vincent P (2000) A neural probabilistic language model. Adv Neural Inf Process Syst 13
45. Gopal S, Patro K, Sahu KK (2015) Normalization: a preprocessing stage. arXiv:1503.06462
46. OECD (2007) Data and Metadata Reporting and Presentation Handbook, p 50. https://www.oecd-ilibrary.org/content/publication/9789264030336-en. https://doi.org/10.1787/9789264030336-en
47. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: WSDM 2015 - proceedings of the 8th ACM international conference on web search and data mining, pp 399–408
48. Koltsova O, Alexeeva S, Pashakhin S, Koltsov S (2020) Polsentilex: sentiment detection in socio-political discussions on Russian social media. In: Artificial intelligence and natural language, 9th conference, AINL, pp 1–16
49. Lago J, Marcjasz G, De Schutter B, Weron R (2021) Forecasting day-ahead electricity prices: a review of state-of-the-art algorithms, best practices and an open-access benchmark. Appl Energy 293:116983. https://doi.org/10.1016/j.apenergy.2021.116983
50. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7(1):1–30
51. Lesche B (1982) Instabilities of Rényi entropies. J Stat Phys 27:419–422

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.