Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2022) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

11th International Young Scientist Conference on Computational Science

An overview of machine learning methods used in forecasting financial time series from news

Grigoryev Mikhail[a,*]

[a]ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia

Abstract

We live in a world filled with media having an impact on almost every aspect of our lives. Thus, a hypothesis can be formulated that news can influence the way we behave economically. This research is an endeavour into the methods of processing text-based information such as news and using them to predict economical variables related to some form of consumption. The overview of those methods fills the knowledge gap between the topics of natural language processing (NLP) and time series forecasting.

Keywords: Machine Learning ; Natural Language Processing ; Topic Analysis ; Autoregressive Models ; ARIMAX

1. Introduction

As methods of machine learning evolve and improve, the spheres of natural language processing (NLP) and time series forecasting grow in popularity among researchers across a range of sciences. However, there is a knowledge gap between the two topics as simultaneous usage of NLP-related and process modeling methods is quite rare in practice.

Consequently, the aim of this research work is to highlight usecases that intersect NLP and time series forecasting in applied sciences. As of today, the majority of people have access to some form of media, be it social media within the realms of the Internet or news presented via television or newspapers. Thus, a valid research niche that combines the methods of NLP and process forecasting is the usage of text-based information (news and other media) as predictors for estimating future values of economical parameters. Such endeavors may give some insight into causality between the media surrounding us and the way we behave financially.

* Corresponding author. Tel.: +7-921-592-4920.
E-mail address: mikegrig@inbox.ru

Due to the specificity of such research, the body of this work is divided into four subsections, the first highlighting the methods of NLP that find use in processing news or similar textual data. Subsections two and three are dedicated to time series forecasting (without and with exogenous variables). Finally, the last subsection combines the knowledge described previously and answers questions, whether it is possible to use text-based information for mentioned purposes.

2. Literature overview

2.1. Topic analysis

The main goal of this part of literature review is to overview the conventional methods of topic analysis that can be used for processing news headlines and similar textual information. The methods mentioned here range from classic LDA to modern BERTopic.

Topic analysis is one of the tasks of natural language processing (NLP). It consists of embedding human-legible texts into machine-readable embeddigs, dimension reduction and further clustering of the embeddings. Such clusters can be interpreted as topics found in the texts, hence the name "topic analysis". One of the classical methodologies of topic analysis is the "bag-of-words", an approach for viewing documents as unstructured vocabularies of words with different frequencies. It was the foundation for a plethora of commonly used topic analysis tools such as LDA, NMF, Top2Vec and CTM.

Authors of the publication [18] extracted topics from scientific literature via custom two-stage pipelines. The first step in every experiment was to embed the text into machine-interpretable numeric vectors. Then in the second step different algorithms were used to cluster the vectors, validate the extracted clusters and output an array of topics.

For embedding scientific texts several approaches including word2vec, POS2vec, word-position2vec and LDA2vec were used individually as well as combined into the improved word embedding based scheme. The latter model takes scientific abstract sentences as input, which is processed by word2vec, a simple embedder that transforms all words within the sentence into vectors without taking the word position into account. Then the authors computed averages of word vectors within each sentence so that the resulting vectors could be used as sentence representations. Before averaging, stop-words have been removed due to this tactic yielding better results. After that, POS2vec (part-of-speech2vec) was used on each word in the sentence to generate POS tags. Then, word-position2vec was employed to generate position vectors for the words and the whole sentence. The latter was concatenated to the vectors obtained earlier. Lastly, the sentences were further processed by LDA2vec. Its output vector was concatenated to yield the final ensemble embedding.

In the publication K-means, K-modes, K-means++, SOM and DIANA were chosen as the clustering algorithms. Classical K-means is a simple partition-based clustering method, while K-modes is its extension better suited for categorical data. As K-means' performance highly depends on the initial random seed for the coordinates of cluster centers, K-means++ uses a heuristic function to determine the optimal initial cluster centers. Self-organizing maps (SOM) method is based on clustering and dimension reduction. Divisive analysis clustering (DIANA) is a divisive hierarchical clustering method. Finally, the authors proposed iterative voting consensus (IVC) – a consensus function which seeks cohesion between different clustering methods on the same data and thus lets those algorithms compensate for each others' weaknesses.

The authors have compared 43 different configurations of topic analysis pipeline on the same dataset of scientific abstracts. The models were compared by values of the following metrics: Jaccard Coefficient (JC), FM and F1 defined by the equations 1, 2, 3, accordingly.

$$JC = \frac{a}{a+b+c} \tag{1}$$

$$FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c}\right)^{0.5} \tag{2}$$

$$F1 = \frac{2a^2}{2a^2 + ac + ab} \tag{3}$$

Here a denotes the number of pairs of instances which are grouped in the same cluster and fall in the same category in the initial dataset. The value b stands for pairs from the same category which were not grouped in the same cluster. Lastly, c denotes pairs from different categories which were grouped together. The higher the metrics described above, the better results models produce.

According to the results presented in the article and in figure 1 ([18]), among single method clustering strategies, LDA (latent Dirichlet allocation) outperformed other algorithms, with DIANA taking the second place. Results of clustering seemed to have improved significantly when integrating LDA2vec embedding scheme with clustering algorithms.

Even better results were shown by the proposed clustering ensemble framework which utilizes IVC for combining the mentioned methods. Thus, such proposed ensemble approach can lead to better topic analysis models.

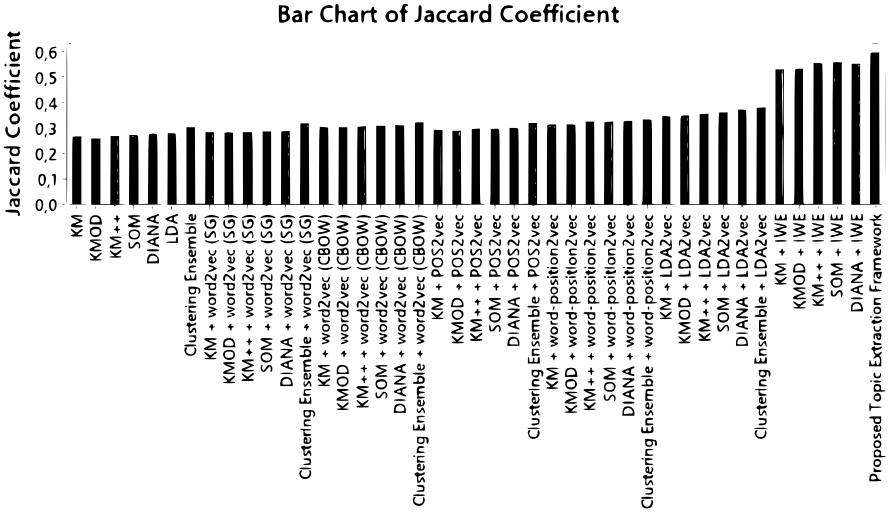

Fig. 1. Jaccard Coefficient metric for different combinations of embedding and clustering methods.

As opposed to the example presented above, topic analysis can solve not only the problem of extracting broad semantic themes from texts but also some more nuanced information. Authors of the article [8] propose their two-step approach for aspect-based sentiment analysis. Their procedure lets evaluate reviews in terms of how positive or negative they are and also define aspects of the review object which are evaluated. In the publication, restaurant and laptop reviews were used as input data. The proposed model outputs a tuple in form of (sentiment, aspect). For example, a review of a restaurant could output (good, ambiance) or (bad, food).

For this finesse task a two-step procedure called JASen was developed. It consists of sentence embedding and classification via convolutional neural networks (CNNs). The former takes labeled data for embedding training and unlabeled data for further classification.

The results shown by the JASen model were compared with those of previously developed approaches. Values of aspect identification and sentiment polarity classification are presented in the tables 1 and 2 below ([8]).

This way, the authors have proposed a model fine-tuned for joint aspect-sentiment extraction which outperformed earlier methods for this task.

Table 1. Quantitative evaluation on aspect identification (%).

Methods	Restaurants				Laptops			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	61.43	50.12	50.26	42.31	53.84	58.79	54.64	52.18
W2VLDA	70.75	58.82	57.44	51.40	64.94	67.78	65.79	63.44
BERT	72.98	58.20	74.63	55.72	67.52	68.26	67.29	65.45
JASen	83.83	64.73	72.95	66.28	71.01	69.55	71.31	69.69

Table 2. Quantitative evaluation on sentiment polarity classification (%).

Methods	Restaurants				Laptops			
	Accuracy	Precision	Recall	macro-F1	Accuracy	Precision	Recall	macro-F1
CosSim	70.14	74.72	61.26	59.89	68.73	69.91	68.95	68.41
W2VLDA	74.32	75.66	70.52	67.23	71.06	71.62	71.37	71.22
BERT	77.48	77.62	73.95	73.82	69.71	70.10	70.26	70.08
JASen	81.96	82.85	78.11	79.44	74.59	74.69	74.65	74.59

However, more often the extraction of topics is enough. In contrast to the ensemble approach proposed in the publication [18], authors of the article [6] leverage different modified algorithms that compose BERTopic – a modern and efficient topic analysis tool.

Its defining feature is the ability to extract topics dynamically and to build topic popularity time series. This is allowed by using c-TF-IDF (class term frequency inverse document frequency) representations of topics defined by equation 4.

$$\text{c\_TF\_IDF} = \underbrace{\frac{n_t}{\sum\limits_{k \in \text{class}} n_k}}_{\text{c-TF}} \cdot \underbrace{\log\left[\frac{|D|}{|\{d_i \in D | t \in d_i\}|}\right]}_{\text{IDF}} \tag{4}$$

Here c-TF is defined by the number of instances of word t divided by the number of words in the class, $|D|$ - number of documents in the dataset, $|\{d_i \in D | t \in d_i\}|$ – number of documents in dataset, containing the word t. First, BERTopic is fitted on the entire dataset to create a global view of topics. Using c-TF-IDF is efficient, as IDF is computed globally and only c-TF needs to be computed at each timestep.

In the publication, newly proposed BERTopic is compared to its predecessors: LDA, NMF, top2vec and CTM by means of topic coherence (TC) and topic diversity (TD) metrics on three datasets of news and twitter posts. According to the article, BERTopic has high coherence scores across all datasets. In terms of topic diversity, it is outperformed by CTM.

One of the major strengths of BERTopic is that it has the capability to operate in the multilingual mode. The default embedding model "all-MiniLM-L6-v2" shows both moderately high topic coherence and diversity across all supported languages in addition to it being lightweight.

More detailed comparison of BERTopic with other methods of topic analysis is presented in the publication [4]. The authors have chosen latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), top2vec and BERTopic in the task of topic analysis on the dataset of Twitter posts.

LDA is a generative probabilistic model for discrete data, which could be viewed as three-level hierarchical Bayesian clustering algorithm. Each document is represented as a mixture of topics with corresponding probabilities and each topic is a mixture over the collection of topic probabilities.

NMF is a decompositional method which works of TF-IDF transformed data by breaking down the input term-document matrix (A) into a pair of lower-ranking matrices:

- terms-topics (W) matrix containing basis vectors;

- topics-documents (H) matrix containing weights.

In NMF all elements in those matrices are non-negative so as to be interpretable.

Top2vec algorithm uses word embeddings via pretrained embedding models so that semantically close words have spatially close embedding vectors. Due to the sparsity of the vector space, a dimension reduction is performed before clustering. Commonly, hierarchical density-based spatial clustering of applications with noise (HDBSCAN) is used to identify dense regions in the reduced vector space of documents. As words that appear in several documents cannot be assigned to one document, they are recognized as noise. Thus, no lemmatization is needed beforehand.

BERTopic uses BERT (bidirectional encoder representations from transformers) pretrained embedders alongside with sentence-transformers for turning documents into vectors across more than 50 languages. Similarly to top2vec, BERTopic uses uniform manifold approximation and projection (UMAP) for dimension reduction and HDBSCAN for clustering. The principal difference between BERTopic and top2vec is that the latter utilizes c-TF-IDF metric instead of normal TF-IDF. Thus, importance of words in clusters and not in documents is taken into account. The usage of HDBSCAN eliminates the need for lemmatization.

Human interpretation of the extracted topics has shown that BERTopic and NMF perform the best among four tested methods. Despite that top2vec and BERTopic both use pretrained embedders for the representation of documents in vector space, many topics extracted by top2vec contained several semantic concepts or intersected each other. LDA gave the least legible results of all methods.

The authors mentioned several drawbacks of BERTopic models such as them yielding large numbers of clusters which leads to the need in manual topic processing. Another mentioned disadvantage is that one document could potentially be assigned to just one topic, which often does not reflect reality. In spite of this, BERTopic still can be characterized as one of the most successful and universally applicable methods of topic analysis due to high model quality and efficiency.

2.2. Time series prediction

Time series forecasting is one of the most important tasks related to time-dependent processes. A plethora of methods for predicting time series have been developed and the most practical ones were described in the publication [19]. The authors group all methods into two major categories: parametric and non-parametric.

Among the parametric methods the following were described:

1. Moving averages models (MA), simple models that view predicted values at a certain timestep as averages over some interval before this step. MA model is defined by the equation 5.

$$z_{t+1} = \frac{\sum_{i=0}^{r-1} z_{t-i}}{r} \tag{5}$$

   Here, t is the timestep, z_{t+1} is the predicted value and r is the number of observations included in the average. Such models are simple but lack quality, when data has seasonality, trend or high-frequency noise.
2. Simple exponential smoothing models (SES) are conceptually close to MA. The main difference is that all previous timesteps are taken into account with different weights, which exponentially decrease the further away from prediction the timestep is. They are defined by the equation 6.

$$z_{t+1} = L_t = \sum_{i=0}^{t-1} \alpha(1-\alpha)^i z_{t-i} \tag{6}$$

Here $\alpha \in (0, 1)$ is the weight for constant smoothing and L_t is the estimate for the next step at time t. For ease of computation, recurrent simplification can be formalized by equation 7.

$$L_t = \alpha z_t + (1 - \alpha) L_{t-1} \tag{7}$$

Initially, it is supposed that $L_1 = z_1$. The main drawback of the method is the difficulty in optimizing α and inaccurate results when dealing with trends.

3. Holt's exponential smoothing (HES) is an extension of SES that utilizes a second smoothing constant β for modeling trends. HES models are defined by equations 8 – 10.

$$L_t = \alpha z_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \tag{8}$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta) T_{t-1} \tag{9}$$

$$z_{t+h} = L_t + hT_t \tag{10}$$

Here, L and T are the level and trend components, accordingly, and h is the prediction horizon (more than 1 step as opposed to SES). Initially, $L_1 = z_1$ and $T_1 = z_2 - z_1$. As in SES method, the most difficult part is the optimization of α and β constants.

4. Holt-Winters' seasonal exponential smoothing models (HW) iterate on HES method principles and additionally deal with seasonality by adding another constant γ. More commonly used additive HW models (AHW) are defined by equations 11 – 14.

$$L_t = \alpha(z_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1}) \tag{11}$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta) T_{t-1} \tag{12}$$

$$S_t = \gamma(z_t - L_t) + (1 - \gamma) S_{t-s} \tag{13}$$

$$z_{t+h} = L_t + hT_t + S_{t-s+h} \tag{14}$$

Here, S is the seasonal component and s denotes the number of timesteps that make a full seasonal cycle. L is usually initialized by the equation 15.

$$L_s = \frac{1}{s} \sum_{i=1}^{s} z_i \tag{15}$$

T with equation 16.

$$T_s = \frac{1}{s} \sum_{i=1}^{s} \frac{z_{s+i} - z_i}{s} \tag{16}$$

and S indexes are computed according to equation 17:

$$S_i = z_i - L_s, \quad i \in [1, s] \tag{17}$$

5. (S)ARIMA, (seasonal) autoregressive integrated moving average models provide even higher quality forecasts for stochastic time series and conceptually include three operations: autoregression AR with parameter p, intergration I with parameter d and moving averages MA with parameter q. Here, integration refers to taking successive differences from the time series ($\Delta z_t = z_t - z_{t-1}$) so as to make the time series stationary, i.e. with no trend and constant deviation. ARIMA of order (p, d, q) is defined by the

equation 18.

$$I'_t = \Delta^d z_t = \delta + \sum_{i=1}^{p} \varphi_i I'_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t \tag{18}$$

Here, I'_t is the value of the d-times differenced stationary time series, φ_i are AR parameters with lags up to p and θ_i are MA parameters with lags up to q. Parameter δ is the initial level of the model and the last term e_t denotes white noise with zero mean and constant deviation. If $d > 1$, the constant δ can be omitted. SARIMA model is an extension of ARIMA that takes in four additional parameters and models seasonality. It is essentially ARIMA (p, d, q) with a seasonal part added. SARIMA of order $(p, d, q) \times (P, D, Q)_s$, where P, D and Q are maximum lags of AR, differencing degree and maximum lags of MA of the seasonal component denoted as I''_t below 19.

$$I_t = \underbrace{\delta + \sum_{i=1}^{p} \varphi_i I'_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t}_{\text{ARIMA}} + \underbrace{\sum_{i=1}^{P} \Phi_i I''_{t-is} + \sum_{i=1}^{Q} \Theta_i e_{t-is}}_{I''_t = \Delta^D z_t} \tag{19}$$

As in ARIMA, here δ can be omitted if $d + D > 1$. (S)ARIMA models are the most commonly used methods of time series forecasting in the modern practice due to algorithm efficiency, ease of use and high quality of predictions on a large range of time series. A detailed usage example of ARIMA on the COVID-19 epidemic dataset is presented in the publication [1]. In the article, as ARIMA is a widely-used model utilized for time series forecasting, no testing of the model was presented. The authors predicted values for the differenced COVID-19 dataset and provided confidence intervals for the forecast. This, as well as the autocorrelation analysis before model fitting gives insight into the real-world usage of ARIMA models on actual data, which provides estimations for the future.

In addition, the authors gave a brief description of non-parametric methods for forecasting processes, which include the following:

1. Artificial neural networks (ANNs) are models comprised of neurons – objects capable of taking several inputs, combining them linearly with a certain bias and feeding the output to an activation function (often sigmoid or relu). The output of a single neuron can be described by the equation 20.

$$f(\text{net}) = f\left(\sum_{i=1}^{l} w_i x_i + b\right) \tag{20}$$

Here, x_i and w_i are inputs and weights, b is a bias and f is the activation function. Arranging neurons in several layers (multiple layer perceptron, MLP) was allowed by introducing backpropagation learning algorithms, where the weights are adjusted to better fit the data in reverse layer order. An example usage of multilayer feed-forward neural networks (MLFFNN) is presented in the publication [11]. The authors showed that even such simple architecture of ANNs can produce good quality forecasts. The paper contains R-metrics (correlation coefficients between predictions and test data) for a dataset of wind speed time series for MLFFNN with values as high as 0.9995. In addition, the authors suggest adaptive neuro-fuzzy inference system (ANFIS) – a method that combines ANNs with fuzzy inference system. The latter implies the usage of fuzzy-parameters (if-else rules manually trained by specialists in the research field). According to the results presented in the article, the utilization of such rules slightly improves time series prediction as opposed to MLFFNN. In the networks mentioned above, neuron-to-neuron signals flow only from input to output. This does not apply to recursive neural networks (RNNs), where neurons form a cycle and the signal has several flow directions. In the simple recurrent network (SRN), the state of a chosen layer within a cycle is conditioned by a context layer on its previous state.

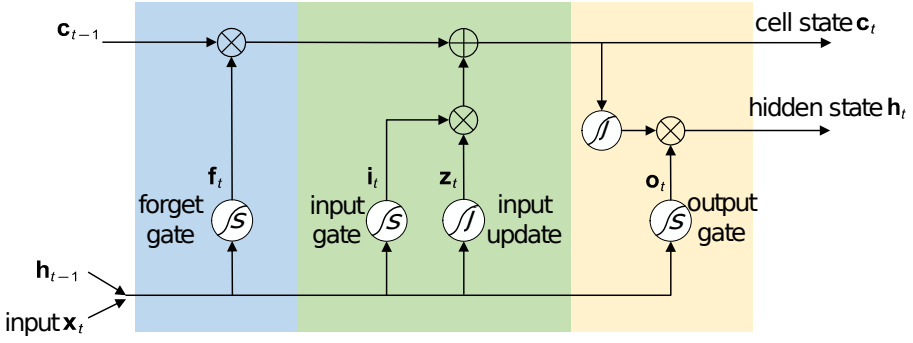

Fig. 2. An illustration of LSTM memory block.

This creates short-term memory, i.e. the ability of the network to store complex signals. The most advanced RNNs for forecasting time series are the long short-term memory (LSTM) networks. LSTMs are described by the authors of the publication [7] as essentially RNNs tuned for learning long-term dependencies via components called “memory blocks”. The latter are recurrent subnets comprised of memory cells and gates. Memory cells’ role is to remember the current state of the network and the gates are in charge of controlling the flow of signal in the network. By their role, gates are classified as input (control the amount of new input data that flows into the network), output (the amount of information that flows out of the memory block) and forget gates (decide the amount of information that remains in the current cell). The layout of a memory block is presented in figure 2 ([7]). It is first decided, which old information stays in the cell, then new information is chosen to be stored. Updating of the cell state is performed by multiplying the old memory state by the forget gate output. Then the element-wise product of the input gate and the candidate values for the new information is added. This procedure lets LSTMs predict time-series with long-term dependencies. Several architectures of LSTM models exist that differ by how the neurons in each memory block are connected. As opposed to fully-connected neurons in the traditional LSTM, they are randomly connected in randomly connected LSTMs (RCLSTM), proposed by the authors of the paper [7]. This architecture more closely resembles real synapses in the human brain, which was one of the reasons for the proposal. The authors tested RCLSTM on several datasets including a dataset of traffic information. The influence of the percentage of connected neurons on the predictive quality was estimated my comparison of RMSE metrics. RCLSTM took less computational resources than standard LSTM, however seemed to be outperformed by it. Taking the above into account, RCLSTM still showed better predictions than commonly used methods such as SVR (support vector machines), ARIMA and FFNN (feed-forward neural networks). This implies that random connection of neurons within LSTMs lets significantly reduce computing time while slightly compromising quality. Even more structurally complicated versions of LSTMs exist. The authors of [15] suggested GeoMAN (”multi-level attention network to predict the the readings of a geo-sensor...”), that is tuned for spatio-temporal correlations. In the paper the authors test this network on a set of geosensors, each providing time-dependent signals (time series). The network utilizes the multi-level attention mechanism including local and global spatial attention as well as temporal attention. This lets the network extract both correlations between different geosensors and also within one time series. Additionally, the authors implemented external factor fusion module, a mechanism of including exogenous variables from different domains for better predictions. The actual architecture of the GeoMAN model is depicted in the figure 3 ([15]).

It can be essentially broken down to two LSTM models, one acting as an encoder for spatial attention and the other as a decoder for temporal. The model was tested on spatio-temporal datasets of water and air quality and within the experiment it was also compared to nine commonly used baseline models including ARIMA, VAR (vector auto-regressive) and many variations of RNNs. According to the publication, on

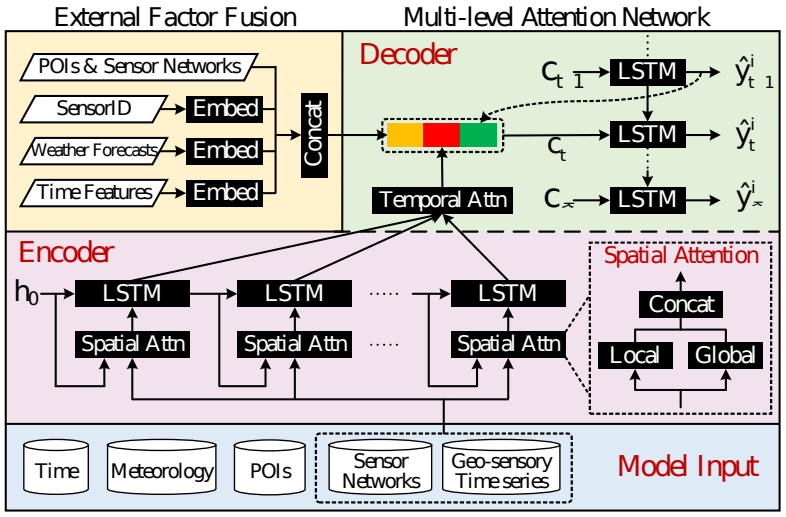

Fig. 3. The framework of GeoMAN.

spatio-temporal data GeoMAN outperforms all baseline models, both parametric and non-parametric, judging by RMSE values. Thus, the paper [15] proves that RNN models can be effectively engineered to perform well on spatio-temporal data.

2. Support vector machines (SVMs) are algorithms conceptually close to ANNs but minimizing the training error while also minimizing the upper bound on the error when the model is applied to test data. The classical illustration to SVMs, binary classification, is presented in figure 4 ([19]).

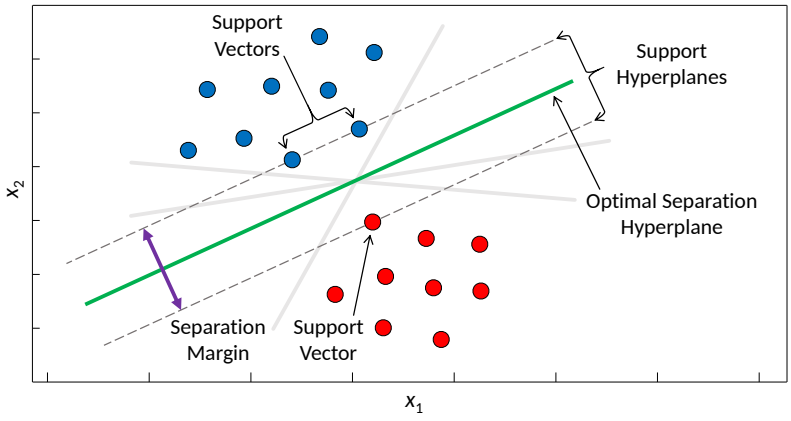

Fig. 4. Binary classification using support vector machines.

The figure shows a set of hyperplanes dividing the two classes. An SVM seeks the optimal divider, where the highest separation margin (distance from divider to support vectors) is reached. As a part of a broader research, SVMs are used along other models in the publication [11] for predicting wind speeds. It showed highly accurate predictions with R-statistics of up to 0.9938, which is accurate, albeit less accurate than the other methods that comprised the research in [11].

3. k-nearest neighbors (kNN) are in nature similarity-based classification algorithms. For forecasting z_{m+1} for a time series $Z = (z_1, \cdots, z_m)$ the algorithm uses last l timesteps as query Q and searches for k most

similar subsequences to Q by sliding a window of size l along the time series. Then, the values of $S_{l+1}^{(j)}$ are averaged in an ensemble to predict z_{m+1} as depicted in equation 21.

$$f(S) = \frac{1}{k} \sum_{j=1}^{k} S_{l+1}^{(j)} \tag{21}$$

Such kNN models have shown the capability to model highly complex non-linear patterns, especially in form of time series prediction with invariance kNN (kNN-TSPI). The later algorithm deals with trivial matches, amplitude, offset and complexity invariance.

4. The authors of the publication [20] offered a hybrid method that combined state space models (SSMs) with deep neural networks. Principally, SSMs utilize a latent space that is used to encode time series' features: level, trend and seasonality. Linear SSMs are described by the equation 22.

$$\mathbf{l}_t = \mathbf{F}_t \mathbf{l}_{t-1} + \mathbf{g}_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1) \tag{22}$$

Here, $\mathbf{l}_t$ is the state of the latent space at timestep t, $\mathbf{F}_t$ is a deterministic transition matrix and $\mathbf{g}_t \varepsilon_t$ is a random innovation. The predicted value is defined by the equation 23.

$$z_t = \mathbf{a}_t^\top \mathbf{l}_{t-1} + b_t + \sigma_t \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1) \tag{23}$$

Here, $\mathbf{a}^\top \in \mathbb{R}^L$, $\sigma_t \in \mathbb{R}^{>0}$ and $b_t \in \mathbb{R}$ are time-varying parameters of the model. SSMs in general are optimized for applications where the time series are structured and this structure is well-known. The major drawback of SSMs is that multivariate analysis of several correlated time series is impossible as the model if fitted onto one process at a time. In the publication, a hybrid model is proposed, that takes advantage of the benefits of both SSMs and RNNs. The model is comprised of an RNN that is used to parametrize a particular linear SSM. RNN's parameters are learned simultaneously from the whole dataset with all additional covariates, which lets the extraction of shared patterns.
This "orchestration" of several SSMs with an RNN is a complicated model architecture, which was quantitatively tested by the authors of the paper [20] against more conventional baseline models such as ARIMA and DeepAR, an RNN-based method. The models were used to forecast electricity consumption and traffic occupancy rates. They were compared by the values of standard $p50$ and $p90$ quantile loss metrics. On the one hand, the majority of the tests with forecasting long-term horizons showed that the proposed hybrid method slightly outperformed the others. On the other hand, DeepAR – RNN – performed better at predicting a short forecast horizon.

2.3. Prediction with exogenous variables

According to [22], "the variables that show differences we wish to explain are called endogenous, while the variables used to explain the differences are called exogenous". If a hypothesis exists, that the observed changes in any variable (called endogenous) are caused by changes in some other variables, the latter are referred to as exogenous.

Victor Gijsbers, the author of the publication [5], elaborates on the concept of causality and it being perceived. According to the paper and despite the arguments againt the Humean regularity theory that views any perception of a cause-effect pair as universal, experiencing causation leaves open the possibility that either the causality itself or its experience depend on some external events. This implies, that at least for statistical data on human behavior, which cannot be considered strongly local, the perception of causality ($a^* \rightarrow b^*$) can theoretically be generalized to ($A \rightarrow B$). Here, a^* and b^* denote specific local observations of random variables A and B, while the arrow depicts causality.

Thus, if the chosen exogenous variables either cause the changes in the predicted variables or have an indirect impact on them, the usage of such variables in a forecasting model can increase its performance. An example of such increase is presented in the publication [24]. The authors tested ARIMA and ARIMAX,

the latter being the same autoregressive model but capable of including exogenous context, in the task of sugarcane yield forecasting in Haryana, India. According to the article, ARIMAX, utilizing exogenous weather data for fitting the model, showed consistently better results than the baseline ARIMA. The values of RMSE, used to compare the models, are given in table 3 ([24]).

Table 3. Comparative view in terms of RMSEs of sugarcane yield forecasts based on ARIMA and ARIMAX models.

District	RMSE	
	ARIMA	ARIMAX
Karnal	7.44	4.18
Ambala	6.07	4.30
Kurukshetra	9.32	6.62

ARIMAX is one of the most widely used models for forecasting with exogenous context, however, many of the methods described in subsection 2.2 allow for the usage of such variables. The authors of the article [2], however did not compare models with exogenous variables to those without them, utilized context information for all of their forecasts across five methods in two variations. This includes:

1. BRNN – bidirectional recurrent neural networks are RNNs that allow both positive and negative time directions. This lets future information (such as future values of exogenous variables) to be included at a certain time frame.
2. CUBIST is rule-based forecasting algorithm that establishes regression models with certain rules based on the input data. General linear regression was used in the paper [2]. Despite the non-linear nature of some time series, implementation of the rules mentioned above lets CUBIST fit complicated temporal patterns.
3. kNN – k-nearest neighbors method described in subsection 2.2 allows for exogenous variables without major modifications as the latter can be used in the Q-queries alongside predicted variables.
4. QRF – quantile random forest is an extension of the random forest (RF) ensemble learning model. In the QRF algorithm, conditional quantiles are utilized. In contrast to conventional RF models, QRF uses full conditional distribution of the predicted variables instead of just the mean.
5. SVR – support vector regression also described in subsection 2.2 similarly to kNN is capable of taking in exogenous variables as-is.

All of the models were used both with and without data preprocessing via variational mode decomposition (VMD). It decomposes a time series into a certain number of mode functions with different sparsities.

The authors of the paper [2] tested the mentioned models against each other in the task of predicting COVID-19 cases in Brazil and USA. Climatic data was used as exogenous variables. According to the results presented in the article, the best predictions for each state in both countries had low values of RRMSE (no more than 6.29% but generally between 1 and 3%). This implies high predictive qualities of models fit with help of exogenous variables, even on large forecast horizons.

Many more variations of autoregressive models that have the capability of including exogenous variables exist. For example, the author of the Master's thesis [3] utilized VARX (vector autoregression with exogenous variables) for predicting stocks and CDS. VARX is an extesion of VAR (vector autoregression) which is closely related to ARIMA. The major benefit of VAR over ARIMA is the capability of multivariate analysis with cross-correlations between different variables. Below, an example equation (24) for VARX with two variables is presented.

$$\begin{pmatrix}\Delta x_t\\ \Delta y_t\end{pmatrix} = \begin{pmatrix}\alpha_0\\ \beta_0\end{pmatrix} + \begin{pmatrix}\sum_{i=1}^{p}\alpha_{1,i} & \sum_{i=1}^{p}\alpha_{2,i}\\ \sum_{i=1}^{p}\beta_{1,i} & \sum_{i=1}^{p}\beta_{2,i}\end{pmatrix}\begin{pmatrix}\Delta x_{t-1}\\ \Delta y_{t-1}\end{pmatrix} + \begin{pmatrix}\alpha_3 & \alpha_4\\ \beta_3 & \beta_4\end{pmatrix}\begin{pmatrix}\Delta z_{1,t-1}\\ \Delta z_{2,t-1}\end{pmatrix} + \begin{pmatrix}\epsilon_t\\ \varepsilon_t\end{pmatrix} \tag{24}$$

Here, x and y are endogenous variables, z_1 and z_2 – exogenous. Coefficients α_1 and β_2 refer to autoregressive coefficients of x and y, accordingly, while α_2 and β_1 are cross-coefficients. The last term in the equation is white noise of the corresponding dimension.

The author used VARX model in the task of predicting CDS and stock prices on the Norwegian market using bond yield and NOK/EUR currency exchange rate (their once-differenced time series) as exogenous variables. According to the thesis, a causal bond between the endogenous variables was found in the way that using exogenous context alongside stock prices as predictors lets accurately forecast CDS, however the opposite is not true. Thus, VARX is a valid method for predicting time series when cross-correlatioins between endogenous variables coexists with important exogenous context.

Another example of an autoregressive model utilizing exogenous variables is brought up in the publication [14]. The authors used nonlinear autoregressive (NARX) model which is a modification of simple AR that allowes for:

1. exogenous variables;
2. nonlinearity in time series patterns.

The model is defined by the equation 25.

$$y_t = f(y_{t-i}, x_{k,t-j}), \qquad \begin{matrix} i \in [1, m] \\ j \in [1, n] \\ k \in [1, n] \end{matrix} \tag{25}$$

Here, y is endogenous, x_k is exogenous and $f(\cdots)$ is a nonlinear function. The authors of the paper used a focused time-delay neural network (FTDNN) for f. The network itself has a complex series-parallel architecture, where the structure of the network changes as the training goes on. According to the article, FTDNN has several advantages over conventionally used RNN including extended capabilities of dealing with short sequence lengths and feature sizes.

The authors tested this NARX model for forecasting traffic flow in the selected portion of a street in the city of Guilin. The model performed reasonably well with R metrics around 0.96 and showed even better predictions after one differencing of the time series. In comparison to more conventional methods, NARX without differencing showed performance better than Holt-Winters method but worse than SARIMA. NARX with one differencing, however, outperformed both models with the R^2 metric reaching the value of 0.957.

2.4. Impact of context information on consumption

This final part of the literature overview is aimed to combine all the methods mentioned in previous subsections for answering the question, whether it is possible to use news as predictors for forecasting economical time series. This raises several questions, each narrower in nature:

1. Do news have enough impact on the consumers to change their behavior?
2. How to process news to be used for fitting models?
3. Which models are capable of predicting consumption or other economical data using news.

The first question was answered in the publications [10] and [21]. Their authors researched how marketplace rumors and corporate news (perceived mostly negatively) are likely to have an impact on consumers. In the article [10] surveys show that almost a third of the marketplace rumors received by a person may be passed along to other people thus targeting a larger demographical group. The paper [21] elaborated the point that negative corporate news about a brand of products are likely to decrease customer loyalty to the brand and the willingness to buy those products. Thus, news can be a powerful predictor for forecasting consumption, even generalizing away from brand perception.

The second question was partially answered in subsection 2.1, however there were more trivial (and less effective) ways of processing news in the scientific sphere, which will be discussed further. The background

for using exogenous context in time series forecasting, necessary for answering the third question, was given in subsection 2.3. Below, some examples of predicting economical time series using news as predictors are given in chronological order, and thus in the order of natural language processing (NLP) advancement.

The authors of [17] utilized simple procedures of sentiment extraction from news headlines. This was done by counting the number of words from the negative financial lexicon defined in the article [16] and the total number of words in the headline. The ratio between the two numbers is used as the predictor for forecasting economical parameters such as Dow Jones Industrial Average, trading volume, volatility (VIX) and the price of gold. The authors used autoregressive models M_0 (no exogenous variables) and M_1 (with exogenous variables) defined by equations 26 and 27.

$$M_0: \quad Y_t = \alpha + \sum_{i=1}^{n} \beta_i Y_{t-i} + \varepsilon_t \tag{26}$$

$$M_1: \quad Y_t = \alpha + \sum_{i=1}^{n} \beta_i Y_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \varepsilon_t \tag{27}$$

Here, Y is endogenous, X is exogenous, α is the level parameter, β_i are autoregressive coefficients, γ_i are exogenous coefficients, ε is noise. The authors simultaneously use the mentioned negative news sentiment (NNS) alongside Twitter investor sentiment (TIS), tweet volume of financial search terms (TV-FST) and daily sentiment index (DSI), processed in similar ways.

The authors showed that the autoregressive model had such good predictive qualities that the addition of exogenous context resulted in just a slight improvement. Nonetheless, there was an increase in forecast quality, which implies some degree of causality between the sentiments found in the news and the observed economical changes.

The authors of the publication [23] went a step further via using embedding models for prediction. The article focuses on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for the task of forecasting stock prices using financial news. As CNNs show generally better results at NLP tasks and RNNs tend to be better at capturing temporal patterns, the authors propose RCNN – the recurrent convolutional neural networks. The architecture of the model consists of four layers:

1. Input layer consisting of the technical indicator layer that takes in a sequence of technical indicators in chronological order and the embedding layer that takes encoded sentences as input.
2. Convolutional layer is composed of convolution, pooling, activation and dropout. In the discussed article, it is tuned for temporal convolution. Thus, this layer can capture local information via combinations of embedded sentences in a window.
3. Recurrent layer is essentially two LSTM models, one for embeddings and one for technical indicators.
4. Output layer is simply a fully-connected softmax activated layer followed by a layer that solves the task of binary classification: $[1, 0]$ for stock price increase and $[0, 1]$ for decrease.

Schematic layout of the RCNN model is presented in figure 5.

According to the results discussed by the authors, the proposed RCNN model outperforms all other baseline models (NNs, RNNs and CNNs) with the exception of EB-CNN (event embedding CNN) which is also a powerful method for modeling content in news articles.

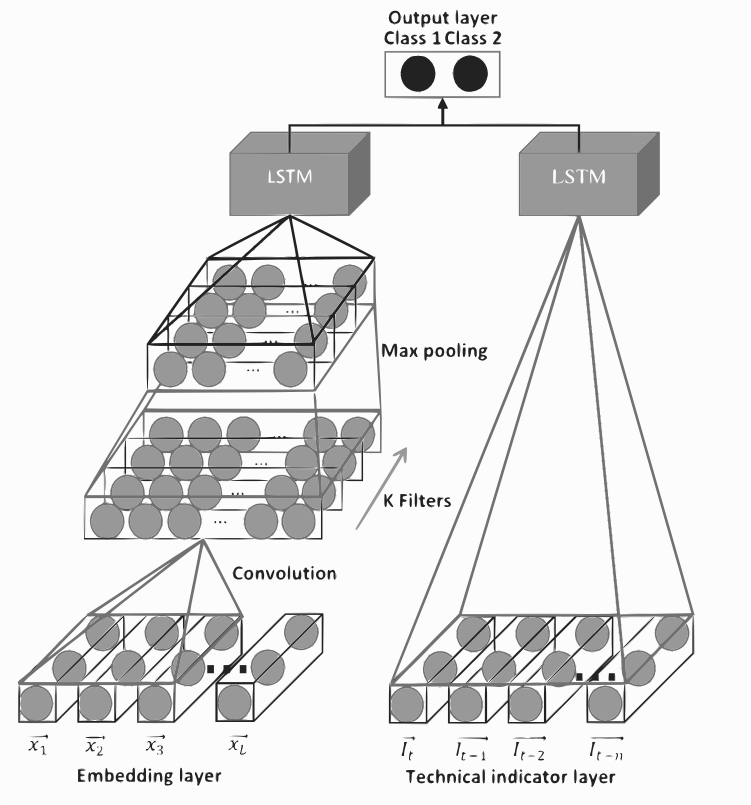

Fig. 5. Architecture of the RCNN model.

A more contemporary method for processing textual information and using it for forecasting time series is described in the publication [12]. The authors researched the ability of news and social media data to predict cryptocurrency prices. They used a modern pipeline consisting of:

- a tokenizer to remove stop characters and punctuation (spaCy was used);
- a vectorizer to embed the tokenized text into numerical data;
- a classifier to learn feature weights (logistic regression and naive Bayes were chosen as the best performing algorithms for the task).

According to the published results, the models extracted general trends for price growth over the examined time period well, however predictions of daily price changes were quite inaccurate. Specifically, predicting fluctuations going against the general trend was a problem. Nonetheless, most of the predictions were correct, which implies the ability of non-technical data such as news to be used as predictors for financial variables.

As the next step in developing a working forecasting model powered by news headlines, the authors of the article [13] have proposed a framework tuned for this task. The designed text-based framework (TBF) takes full advantage of the textual input data and predicts agricultural futures such as soybean prices. The structure of TBF is depicted in figure 6.

The framework combines tasks of topic analysis described in subsection 2.1, sentiment analysis and time series forecasting. The most intriguing part of the framework is in the steps 2 and 3. The model identifies influential factors if those are easily quantifiable or performs sentiment analysis instead otherwise. This lets all the news to be utilized in model fitting without losing quality by including vague news.

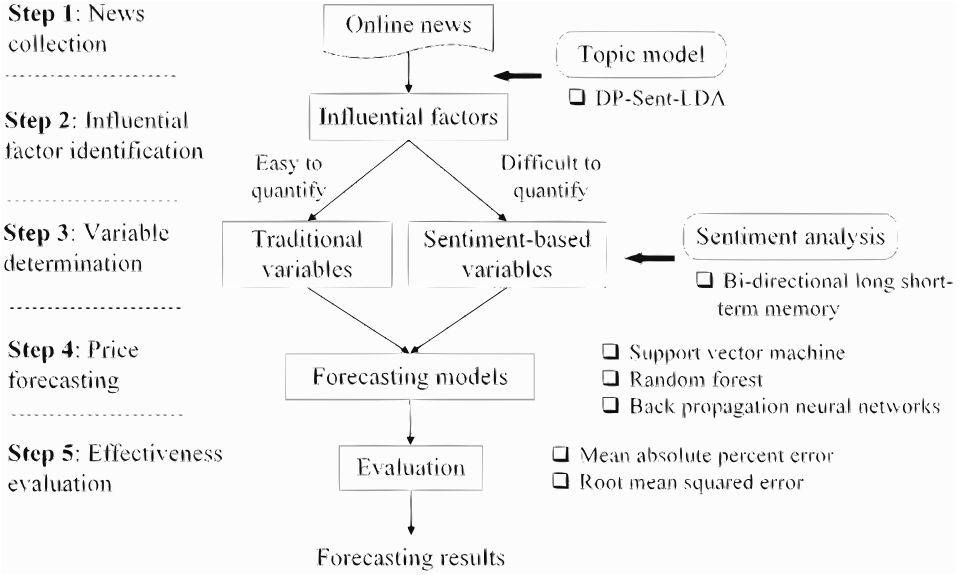

Fig. 6. Structure of the text-based framework.

For the forecasting section of the framework, the authors used SVR (support vector regression), RF (random forest) and BPNN (back propagation neural network). The authors compared all TBF-based models to baseline ARIMA. According to the results presented in the article, TBF-RF and TBF-SVR performed as well as ARIMA with small prediction horizon while TBF-BPNN had significantly higher forecasting errors. Interestingly, starting from lags of 40-50 days, prediction error of ARIMA starts growing linearly with the lag. TBF models' errors reach a plateau after 10-15 days and remain stable regardless how large is the time lag. Thus, in medium- and long-term predictions all of the TBF-based models outperform ARIMA.

Less task-specific methods were used by the author of the PhD thesis [9] in the task of Apple stock price forecasting. An array of conventional time series forecasting models allowing for exogenous context was used in conjunction with NLP methods for processing news headlines. For news sentiment analysis the following models were used:

- LSTM (long short-term memory networks), a type of RNN discussed in subsection 2.2.
- GRU (gated recurrent unit) is a modified LSTM, where the addition of new information and retention of old information are interdependent.
- BERT (bidirectional encoder representation from transformers) is one of the cutting-edge NLP models with complicated architecture that takes advantage of transformers – RNNs tuned for sequential information and NLP tasks.

For time series prediction the following methods were utilized:

- (S)ARIMA(X), (seasonal) autoregressive integrated moving averages (with exogenous predictors) – classical time series forecasting models discussed in subsection 2.2.
- Linear regression using PCA (principal component analysis), a method that performs dimension reduction and viewing the dependent variable as a linear combination of the predictors in the new latent space.
- RF (random forest method) is an ensemble learning method that consists of multiple decision trees. Every tree produces its prediction and they all are averaged for the total forecast.
- LSTM, described before.

According to the results presented in the article, SARIMAX failed to provide any accurate predictions in this task, while LSTM and, interestingly, linear regression with PCA had low MSE values of forecasts. Poor performance of SARIMAX could be explained by the lack of detail in the input data.

3. Conclusion

This research work highlighted the usage of news and similar text-based contextual information for researching the influence of news on consumer behavior by predicting financial time series. In doing so, NLP methods of topic analysis as well as a myriad of process forecasting methods that either allow exogenous variables or are especially tuned for text-based input, were structurally described and compared to each other in terms of implementation complexity and performance across a range of prediction tasks.

Fortunately, in recent years more and more frameworks are developed that combine NLP and time series forecasting, which gradually closes the mentioned knowledge gap. This scientific work is a contribution into such research acting as a bird's eye view of the most acknowledged methods that might find more use in other practical works.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Anton Kovantsev who supported the choice of the topic and made this research possible. His assistance throughout the process of conducting this research helped it become more logically structured and helpful for the audience.

I would also like to give special thanks to my family and close friends for their continuous support and understanding.

References

[1] Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., Ciccozzi, M., 2020. Application of the arima model on the covid-2019 epidemic dataset. Data in brief 29, 105340.

[2] Da Silva, R.G., Ribeiro, M.H.D.M., Mariani, V.C., dos Santos Coelho, L., 2020. Forecasting brazilian and american covid-19 cases based on artificial intelligence coupled with climatic exogenous variables. Chaos, Solitons & Fractals 139, 110027.

[3] Ding, R., 2021. Empirical Evidence of Lead-Lag Relation between the Norwegian CDS and Stock Markets: Using Vector Autoregression with Exogenous Variables (VARX) and Structured Regularization for Large Vector Autoregressions with Exogenous Variables (VARX-L) Framework. Master's thesis.

[4] Egger, R., Yu, J., 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in Sociology 7.

[5] Gijsbers, V., 2021. Perceiving causation and causal singularism. Synthese 199, 14881–14895.

[6] Grootendorst, M., 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 .

[7] Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., Zhang, H., 2019. Deep learning with long short-term memory for time series prediction. IEEE Communications Magazine 57, 114–119.

[8] Huang, J., Meng, Y., Guo, F., Ji, H., Han, J., 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. arXiv preprint arXiv:2010.06705 .

[9] Jeong, J., 2022. Predicting Apple Stock Price Using News Headlines and Other Features With Classical Time Series Models, Supervised Models, and Machine Learning Models. Ph.D. thesis. UCLA.

[10] Kamins, M.A., Folkes, V.S., Perner, L., 1997. Consumer responses to rumors: Good news, bad news. Journal of consumer psychology 6, 165–187.

[11] Khosravi, A., Machado, L., Nunes, R., 2018. Time-series prediction of wind speed using machine learning algorithms: A case study osorio wind farm, brazil. Applied Energy 224, 550–566.

[12] Lamon, C., Nielsen, E., Redondo, E., 2017. Cryptocurrency price prediction using news and social media sentiment. SMU Data Sci. Rev 1, 1–22.

[13] Li, J., Li, G., Liu, M., Zhu, X., Wei, L., 2022a. A novel text-based framework for forecasting agricultural futures using massive online news headlines. International Journal of Forecasting 38, 35–50.

[14] Li, J., Li, W., Lian, G., 2022b. A nonlinear autoregressive model with exogenous variables for traffic flow forecasting in smaller urban regions. Promet 34, 943–957.

[15] Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y., 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction., in: IJCAI, pp. 3428–3434.
[16] Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of finance 66, 35–65.
[17] Mao, H., Counts, S., Bollen, J., 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051 .
[18] Onan, A., 2019. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. IEEE Access 7, 145614–145633.
[19] Parmezan, A.R.S., Souza, V.M., Batista, G.E., 2019. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. Information sciences 484, 302–337.
[20] Rangapuram, S.S., Seeger, M.W., Gasthaus, J., Stella, L., Wang, Y., Januschowski, T., 2018. Deep state space models for time series forecasting. Advances in neural information processing systems 31.
[21] Sago, B., Hinnenkamp, C., 2014. The impact of significant negative news on consumer behavior towards favorite brands. Global Journal of Business Research 8, 65–72.
[22] Smelser, N.J., Baltes, P.B., et al., 2001. International encyclopedia of the social & behavioral sciences. volume 11. Elsevier Amsterdam.
[23] Vargas, M.R., De Lima, B.S., Evsukoff, A.G., 2017. Deep learning for stock market prediction from financial news articles, in: 2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA), IEEE. pp. 60–65.
[24] VERMA, U., et al., 2022. Arima and arimax models for sugarcane yield forecasting in northern agro-climatic zone of haryana. Journal of Agrometeorology 24, 200–202.