Research article

DOI: https://doi.org/10.48554/SDEE.2024.2.1

An Algorithm for Forecasting Future Trends

Aleksandr Borzov[1*], Victor Sorokin[2], Vitaly Mityazov[2], Daria Shimanova[2]

[1]St. Petersburg Restoration and Construction Institute, St. Petersburg, Russian Federation, priem@spbrsi.ru
[2]Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation, vitya.soroka.02@mail.ru, vitalikmit@mail.ru, shimanova_da@spbstu.ru
[*]Corresponding author: priem@spbrsi.ru

Abstract

The contemporary information landscape is characterised by a huge amount of data available for analysis using a variety of research tools and methods. Considering the limitations of using individual models and methods, it is worth employing an approach that combines functional and logical autoregression methods to conduct a more accurate analysis of trends and topics in the information space. Considering this context, this work aims to develop an algorithm to identify and analyse topics that would be relevant in the future using autoregression methods. The process begins with the quantification and normalisation of data, which significantly affect the quality of analysis. The main focus of this study is to implement the autoregression method to analyse long-term trends and predict future developments in the selected data. The proposed algorithm evaluates the forecast of these future developments and analyses graphical trends, thus conducting a more detailed study and modelling of future data dynamics. The regression coefficient is used as a quality criterion. The algorithm concludes with a polynomial function to help identify topics that will be relevant in the future. Overall, the proposed algorithm can be considered an effective tool for analysing and predicting future trends based on the analysis of historical data, thus contributing to the identification of prospects for technological development.

Keywords: data quantification, time series normalisation, forecasting, autoregression models, data normalization, trends

Citation: Borzov, A., Sorokin, V., Mityazov, V., Shimanova, D., 2024. An Algorithm for Forecasting Future Trends. Sustainable Development and Engineering Economics 2, 1. https://doi.org/10.48554/SDEE.2024.2.1

Научная статья

УДК 330.4

DOI: https://doi.org/10.48554/SDEE.2024.2.1

Разработка Алгоритма Прогнозирования Будущих Тенденций

Александр Борзов[1*], Виктор Сорокин[2], Виталий Митязов[2], Дарья Шиманова[2]

[1]Санкт-Петербургский реставрационно-строительный институт, Санкт-Петербург, Российская Федерация, priem@spbrsi.ru
[2]Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация, vitya.soroka.02@mail.ru, vitalikmit@mail.ru, shimanova_da@spbstu.ru
[*]Автор, ответственный за переписку: priem@spbrsi.ru

Аннотация

Современная информационная среда содержит огромное количество данных, доступных для анализа с использованием множества инструментов и методов исследования. Применение отдельных моделей и методов может быть ограничено, поэтому стоит использовать комбинированный подход, объединяющий методы функциональной и логической авторегрессии для более точного анализа трендов и тематик в информационной среде. Основная цель работы - разработать алгоритм для выявления и анализа будущих актуальных тематик с использованием методов авторегрессии. В работе был разработан эффективный алгоритм для выявления значимых тем в будущем. Процесс начинается с квантификации и нормализации данных, что существенно влияет на качество анализа. Основное внимание уделяется использованию метода авторегрессии для анализа долгосрочных тенденций и прогнозирования будущего развития данных. Алгоритм оценивает прогноз будущего развития и проводит анализ графических трендов для более детального изучения и моделирования будущей динамики данных. Коэффициент регрессии используется в качестве критерия качества, а завершением алгоритма является нахождение полиномиальной функции, что помогает выявить будущие актуальные темы для исследований. В целом, полученный алгоритм представляет собой эффективный инструмент для анализа и прогнозирования будущих тенденций на основе анализа исторических данных, способствуя выявлению перспектив развития технологий.

Ключевые слова: квантификация данных, нормализация временных рядов, прогнозирование, модели авторегрессии, нормализация данных, тенденции.

Цитирование: Борзов А., Сорокин, В., Митязов, В., Шиманова, Д., 2024. Разработка Алгоритма Прогнозирования Будущих Тенденций. Sustainable Development and Engineering Economics 2, 1. https://doi.org/10.48554/SDEE.2024.2.1

1. Introduction

The rapid growth in the volume and diversity of data has increased the importance of data analysis and interpretation. In this context, forecasting of future trends is gradually becoming a key aspect for various industries, such as the marketing, finance and technology sectors. Notably, this process involves analysing current data and applying various methods to predict future events.

In this paper, we develop an algorithm based on functional and logical autoregression methods that allows for the identification and analysis of topics that are likely to be relevant in the future. One of the main steps involved in this algorithm is the quantification and normalisation of data, which is critical to ensure high accuracy of analysis. Furthermore, the widespread use of autoregressive models in time series analysis serves as the basis for identifying long-term trends and cyclical patterns in data, leading to more stable and reliable forecasts. This work is particularly aimed at analysing graphical trends, thus contributing not only to detailing the dynamics under study but also to optimising the process of making strategic decisions based on available data.

2. Literature Review

The information space is loaded with huge amounts of data available for analysis and interpretation using a variety of tools and methods. Therefore, in recent decades, data analysis has emerged as an object of active study, especially in the context of big data and its features. The main focus in this field includes the collection, processing and interpretation of large amounts of information, which requires researchers to have a deep understanding of the methods and tools that contribute to effective analysis.

An important aspect of forecasting is the use of autoregressive models for time series analysis. These models make it possible to identify time dependencies and predict future values based on available data. One such model is functional autoregression, which calculates the current value of a series in terms of its previous values. Notably, this model accounts for both linear and nonlinear dependencies between successive observations, which makes it a powerful tool for time analysis. The fact that functional autoregression is being used in various fields, such as economics, finance, climatology and medicine, confirms its effectiveness and flexibility in the analysis of complex time dependencies (Puchkov and Belyavsky, 2018; Mestre et al., 2021). In contrast to functional autoregression, logical autoregression helps to analyse dependencies between categorical variables in a data sequence. This method allows for the prediction of the next value of a categorical variable based on historical data, thus modelling the probability of each category based on previous values (Huang et al., 2012).

One of the key stages of data analysis for forecasting is data normalization. This process of standardizing variable values not only improves model performance and simulation quality but also ensures consistent and accurate interpretation of results. As already noted by several researchers (Singh and Singh, 2020; Mahmoud et al., 2023), data normalisation is the basis for improving the quality of predictive models because it facilitates fair comparison between variables, which is especially significant when working with large amounts of data.

Although the application of individual models and methods may yield the desired results depending on the context and objectives of a study, one cannot be entirely certain that the obtained results fully reflect the properties of the aspect being studied. To address this, it is necessary to develop algorithms that combine suitable methods and models to create a synergistic effect that enables the in-depth interpretation of results.

Considering the context of this study, the combined use of functional and logical autoregression methods can help to not only analyse the dependencies between variables over time but also identify the relationships between different topics, thus contributing to the detection of significant trends and patterns in the data flow (Huang et al., 2017; Savzikhanova, 2023). Therefore, the framework of this study adopts the combined use of functional and logical autoregression methods to accurately determine trending topics while also analysing the properties of both methods in the context of long-term forecasting.

3. Materials and Methods

3.1 Data quantification

In the initial stage of the proposed algorithm, the specifics of the data submitted as input, initially in text format, are first taken into consideration. Notably, the choice of data source depends on the context and goals of the study. For instance, the source could be the media, social media, web pages, reports, etc. (Di et al., 2017).

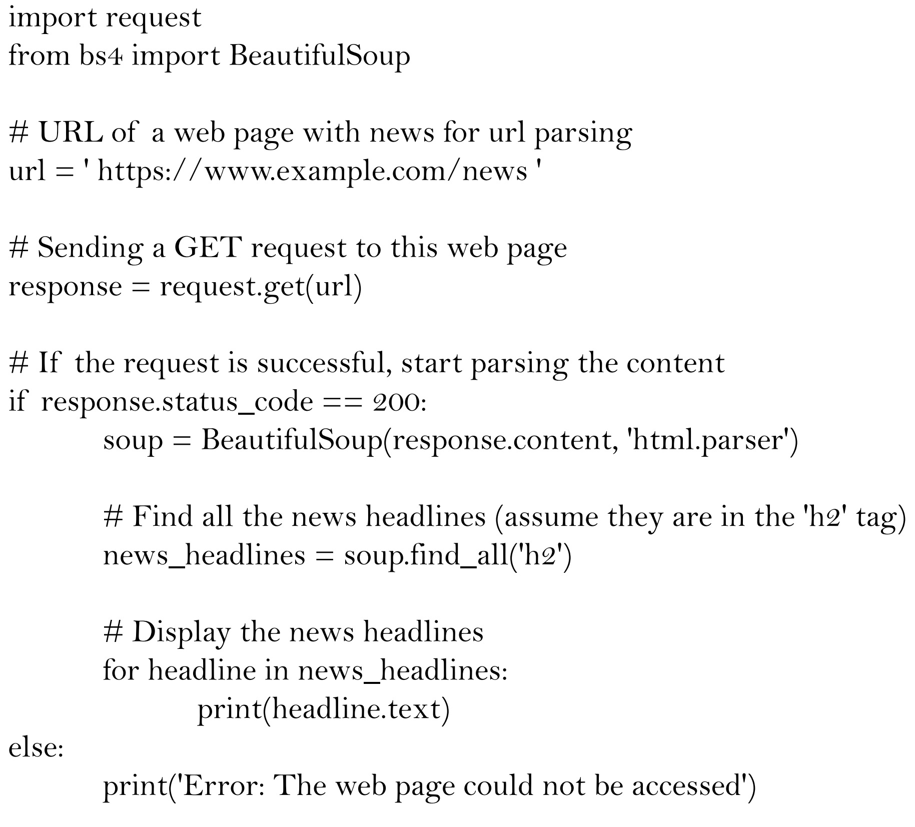
The following are some tools that can be employed to collect text data from various sources:

- Web scraping allows the direct extraction of data from websites from the HTML codes of web pages. For this purpose, Python libraries, such as BeautifulSoup or Scrapy, can be used.

- Many online platforms and services provide application programming interface (API) for accessing their data. The API programming interface allows one to directly receive information from sources such as social networks, news portals, etc.

- RSS feeds, which may be available for some news sites and blogs, make it possible to receive updates automatically in the form of text data.

Figure 1 presents an example of a Python code using the BeautifulSoup library for parsing news from a web page.



```
import request
from bs4 import BeautifulSoup

# URL of  a web page with news for url parsing
url = ' https://www.example.com/news '

# Sending a GET request to this web page
response = request.get(url)

# If  the request is successful, start parsing the content
if  response.status_code == 200:
        soup = BeautifulSoup(response.content, 'html.parser')

        # Find all the news headlines (assume they are in the 'h2' tag)
        news_headlines = soup.find_all('h2')

        # Display the news headlines
        for headline in news_headlines:
                print(headline.text)
else:
        print('Error: The web page could not be accessed')
```

Figure 1. Using the BeautifulSoup library to extract data from a web page.

After the necessary data are obtained, they must be quantified by translating the textual information into structured quantitative indicators (Trewartha et al., 2022; Zhang et al., 2016). For this purpose, text analysis methods can be used. Text analysis allows the extraction of key concepts, themes, emotions and other aspects necessary for research from texts to ultimately present them in the form of digital data (Cover and Thomas, 2005).

In such analyses, texts need to be split so that each word in a single text represents a token. This process is referred to as tokenisation. With regard to the mathematical description of this process (Friedman, 2023; Minogue et al., 2015), assuming that text T has to be tokenised, the text can be represented as a sequence of characters, as follows:

$$T = c_1, c_2, \ldots, c_n \tag{1}$$

Where:

c_i - the i-th item in the text

After the tokenisation process, each word can be represented as a token:

$$T = w_1, w_2, \ldots, w_m \tag{2}$$

Where w_i - the i-th word in the text

Therefore, the tokenisation process can be described in terms of the following function:

$$tokenize(T) = [w_1, w_2, \ldots, w_m] \tag{3}$$

This function converts the input text T into an array of tokens. For example, after tokenization, the sentence “Text tokenization process” will be presented in the form of an array as follows:

$$tokenize(\text{The process of text tokenization}) = [\text{The, process, of, text, tokenization}]$$

In this context, it is worth noting that punctuation marks, numbers, abbreviations and other special characters are also accounted for in the tokenisation process to ensure accurate separation of the text into tokens.

Specialised libraries and tools are often used to divide text into tokens of software code. For example, in Python, the NLTK library can be employed to tokenise text by importing word_tokenize from nltk.tokenize (Zhang et al., 2024; Kadiev and Kadiev, 2016). Figure 2 presents an example demonstrating the use of this library.

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word tokenize

# Input text for tokenization
text = "Example of  the tokenization process in Python."

# Tokenization of  text
tokens = word_tokenize(text, language="russian")

# Tokens output
print(tokens)
```

Figure 2. An example of using the NLTK library.

Therefore, after tokenization, the following list is obtained as output: ‘Example’, ‘process’, ‘tokenisation’, ‘in’, ‘Python’ and ‘.’.

After dividing the text into tokens, the lemmatisation process begins. Lemmatisation involves the transformation of words (tokens) into their basic normalised forms (lemma) using certain rules and algorithms while also accounting for contextual information (Boban et al., 2020; Ozturkmenoglu and

Alpkocak, 2012; Toporkov and Agerri, 2023). Mathematically, the process can be represented as a function applied to each word in the text, as follows:

$$\text{lemma}\left(\text{w}\right) = \text{base\_form} \tag{4}$$

Where:

base_form - the basic form of the word w.

Notably, lemmatisers implement rules and algorithms that account for the morphological characteristics of words. For example, consider the following algorithm comprising a set of conditions and operations for lemmatising verbs into their initial forms:

$\text{lemma}\left(\text{w}\right) =$

$w[:-2]$ if w ends with "-ла",

$w[:-2]$ if w ends with "-ли",

$w[:-2]$ if w ends with "-ть",

otherwise, the word will remain unchanged.

Where:

- w is the root word for lemmatisation,

- $w[:-2]$ indicates removing the last two characters (in this case, the ending) of the word to attain its basic form.

The next step involves clustering the received words based on specific topics. In this context, the main method that can be implemented is TF-IDF. Specifically, this method helps identify keywords or terms that are most characteristic of an individual document in the context of the texts contained in it. The operating principle of TF-IDF (Aizawa, 2003; Dey and Das, 2023; Zhang et al., 2019) involves calculating a measure that evaluates how often a word appears in a document (TF or term frequency). Therefore, the formula for calculating the TF of the word t in document d can be expressed as follows:

$$TF\left(t,d\right) = \frac{n\left(t\right)}{m\left(d\right)} \tag{5}$$

Where:

$n\left(t\right)$ − number of times the word t appears in the document,

$m\left(d\right)$ − total number of words in document d.

This part of the algorithm allows for the evaluation of the importance of a word in a single document.

Next, the inverse document frequency (IDF), which evaluates the uniqueness of a word within a collection of documents, is calculated (Choi and Lee, 2020; Dagdelen et al., 2024). The formula for calculating the IDF for word t in document collection D can be expressed as follows:

$$IDF\left(t,D\right) = \log\left(\frac{N\left(D\right)}{df\left(t\right)}\right) \tag{6}$$

Where:

$N\left(D\right)$ − total number of documents in collection D,

$df\left(t\right)$ − number of documents in the collection in which the word t appears.

Notably, the inverse frequency of a document is output in logarithmic form to reduce the influence of common words and to increase the weight of unique words.

Finally, TF-IDF for the word t in document d can be calculated as the product of TF and IDF as follows:

$$TF - IDF\left(t,d,D\right) = TF\left(t,d\right) * IDF\left(t,D\right) \tag{7}$$

Simply put, words with a high TF-IDF value are those that occur frequently within a single document but are rare when considering the entire collection of texts.

Following this, the latent Dirichlet allocation (LDA) model – a probabilistic model used to analyse the thematic structure of documents – is applied. It can be expressed as follows (Gandhi et al., 2008; Jelodar et al., 2019):

$$p\left(w,z,\theta,\phi|\alpha,\beta\right) = \Pi\left(d=1\right)^{(D)} p\left(\theta_d|\alpha\right)\left(\Pi\left(n=1\right)^{(N)} p\left(z_{(d,n)}|\theta_d\right) p\left(w_{(d,n)}|\phi_{\left(z_{(d,n)}\right)}\right)\right) \tag{8}$$

Where:

D - number of documents,

N - number of words in the document,

K - number of topics,

w - a specific word in the document,

α - Dirichlet hyperparameter for the distribution of topics across documents,

β - Dirichlet hyperparameter for the distribution of words by topic,

θ - distribution of topics in the document,

ϕ - distribution of words for a topic,

z - hidden variable indicating the topic of the word w.

The generative process of this model can be described as follows:

- For each document d in the collection:

- Select the distribution of topics θ_d from Dirichlet (α).

- For each word $w_{(d,n)}$ in the document:

- Choose a topic $z_{(d,n)}$ from multinomial (θ_d),

- Choose a word $w_{(d,n)}$ from multinomial ($\phi_{\left(z_{(d,n)}\right)}$).

The main goal of LDA is to find matrices θ and ϕ for the hyperparameters α and β so as to maximise the plausibility of the data. Mathematically, it boils down to finding the posteriori distributions $p\left(\theta|w,\alpha,\beta\right)$ and $p\left(\phi|w,\alpha,\beta\right)$. In this context, it is also worth noting that a variational method can be employed in the LDA process to maximise the likelihood function, which includes recalculating the parameters and updating the distributions θ and ϕ. Thus, the LDA mathematical model is based on calculating the probability distributions of topics and words, providing an analysis of the thematic structure of documents and highlighting significant topics mentioned in the text data.

After selecting the necessary clusters, the probabilities of each text belonging to a specific topic are generated. Based on the data obtained, one or more classification model are trained (for example, random forest). Subsequently, the trained model was employed to predict the probabilities of each text belonging to the selected clusters. Notably, these probabilities are presented numerically, reflecting the degree of confidence of the model with regard to the text belonging to each cluster selected earlier. In addition, probabilities can be interpreted as the proportion of the presence of a topic in each text.

Thus, the above-mentioned process of quantifying textual data using a probabilistic classification model makes it possible to effectively analyse the text, identify thematic affiliations and make informed decisions based on the results of classification and the probabilities of belonging to clusters.

3.2 Averaging and normalising the received data

In the case of a large dataset, it can be averaged based on a chosen criterion. For example, it can be considered a temporary variable. This is a key action for several reasons:

- Large datasets often contain temporary or random noise, which can make it difficult to identify general trends. Averaging allows for smoothing out these fluctuations, ultimately highlighting more stable and general patterns.

- If the main purpose of a study is to predict a particular trend, averaging can make the data more predictable, as well as improve the predictive ability of models and algorithms.

- Since average values reflect general trends and allow a better understanding of the main characteristics of the data, they help improve the level of interpretation and visualisation.

In this context, the role and relevance of data normalisation, as well as its benefits for analysis and forecasting, must also be emphasised. In the context of previously obtained data being fed into functional and logical autoregression models, the relevance of normalisation becomes even more significant. Considering this aspect, it is worth highlighting the following points:

- When using the operation, more stable and reliable predictive models are created, since the noise and fluctuations in the data are smoothed out, which in turn increases the stability and efficiency of autoregression models.

- Despite changes in the values themselves, normalisation helps preserve the ratios and patterns of the data, which in turn helps identify and retain long-term trends, as well as the cyclicity of autoregressive models.

One of the common methods used for data normalisation is Z-normalisation (standardisation), which modifies the data so that the average value is 0 and the standard deviation is 1. This process can be formulated as follows:

$$z = \left(x - \mu\right) / \left(\sigma\right) \tag{9}$$

Where:

z - the normalised value,

x - initial value of the variable,

μ - average value of the variable,

σ - standard deviation of the variable.

Another such method is min-max normalisation, which scales the data so that it remains within a certain range – often between 0 and 1.

$$x_{norm} = \left(x - x_{min}\right) / \left(x_{max} - x_{min}\right) \tag{10}$$

Where:

x_{norm} - the normalised value,

x - initial value of the variable,

x_{min} - minimum value of the variable,

x_{max} - maximum value of the variable.

The above formula brings the data within the range of 0 to 1, while also accounting for their minimum and maximum values, thus simplifying the comparison and use of features with different ranges of values.

Therefore, it may be concluded that the process of data normalisation brings variables to a common scale, thus standardising their values and preventing distortions in assessing the importance of features. Overall, in the context of data processing, normalisation methods, such as Z-normalisation or min-max normalisation, equip predictive models with the properties of stability, efficiency and greater interpretability of results (Shantal et al., 2023; Peng et al., 2005). By standardising the values of all variables, normalisation offers an opportunity to objectively compare and utilise data in various models and tasks.

4. Results

The application of an autoregression model can be carried out using various methods, such as logistic regression, Markov models or other machine learning methods capable of working with categorical data. In the case of this study, the regression decision tree model was chosen. Notably, in the context of logical autoregression, when the logic is based on a regression decision tree, it indicates that the model is to be used for predicting categorical or qualitative variables based on historical data. The operating principle of a regression decision tree is that the tree is built using nodes, which represent the feature-based data divisions, with the leaves predicting the categorical value. When training this model, the data are first divided into parts, after which the most appropriate categorical value for each division is calculated. The advantages of using a regression decision tree in logical autoregression include ease of interpretation of the rules obtained, the ability to model complex nonlinear dependencies between variables and good generalisation ability.

To predict trends in information flow, the proposed algorithm had to be slightly modified. In particular, the categorical values represent the probabilities of the text belonging to the relevant topic at a specific moment. For example, the trend in the magnitude of the presence of a cluster in texts over a given time period can take the form of a polynomial of the second degree, reflecting the specifics of the dynamics of the selected topic's relevance. In this context, the most relevant example is the history of the development of neural networks. Although the concept of neural networks originated in the 1940s and 1960s, the lack of necessary computing power led to it losing its relevance. However, with the discovery of necessary tools, powerful graphics processors and access to large amounts of data, this niche has regained its relevance. Therefore, on a graph, the dynamics of the urgency of this topic will appear as a polynomial of the second degree. Drawing on this, in the context of researching different niches, it is necessary to look for topics that correspond to the specifics described for the development of neural networks. Furthermore, when using traditional autoregression, it is necessary to predict the values of the selected "trend" topics for the period relevant to the context of the study.

Considering these conditions, the principle for training the model had to be changed – it had to be trained on historical data, after which, based on the values of "relevant topics" obtained using the autoregression method, the probability values of the presence of one or more "potentially relevant" topic should be predicted. After obtaining the predicted values of the share of a particular topic within the average values, a trend can be built, characterising the topic as potentially “relevant” or “irrelevant”.

Although there are several approaches for building a trend, the simplest is to build a linear trend – a straight line reflecting the general trend of the changes in data over time. Linear regression can be conducted to model a linear trend, with time stamps considered as an independent variable and time series values as the dependent variable. Mathematically, the process of obtaining a linear trend model through linear regression can be expressed as follows:

$$Y = a + bt + e \tag{11}$$

Where:

Y - values of the time series;

a - point of intersection with the Y axis, representing the initial value;

b - slope of the linear trend, representing the rate of change;

t - number of intervals, which may have indices of time moments;

e - a random error.

Additionally, to construct a trend line, the logarithmic method can be employed, assuming that the relationship between variables is logarithmic, as follows:

$$ln(Y) = a + b * ln(t) + e \tag{12}$$

Another method for establishing a trend is polynomial regression, which utilises a polynomial to represent the trend. Notably, this technique assists in identifying nonlinear relationships in the data. In this context, it is crucial to determine the appropriate degree of polynomial to ensure an accurate fit without underestimating or oversimplifying the model. For example, consider the following formula:

$$Y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + e \tag{13}$$

Where:

Y - dependent variable (time series value),

x - independent variable (time stamps),

$a_0 + a_0x + a_0x_2 + ... + a_nx^n + e$ - coefficients of the polynomial,

n - order of the polynomial,

e - a random error.

To select the optimal coefficients of the model and determine the order of the polynomial, the least squares method may be employed since it minimises the sum of squares of the difference between the actual and predicted values. In this study, we compared the predicted probabilities of a text belonging to a cluster with the actual trend values.

The main quality criterion considered in this study was the regression coefficient obtained after constructing a trend line. Notably, this coefficient can also be used to assess the quality of the models. Although the other quality parameters, such as the average quadratic error, average absolute error, and coefficient of determination, are important (R^2), they were considered secondary, since the main purpose of the proposed algorithm was to determine the trend of a topic.

In this regard, the specifics of interpreting the regression coefficient must also be discussed. In the case of a positive regression coefficient, the dependent variable increases over time, indicating an upward trend in the data. This property is significant because it allows for a comprehensive understanding of how the dependent variable changes and helps predict future trends. Notably, trends can take various forms – flat sections, oscillations or polynomial curves. By analysing data using a positive regression coefficient, we can identify patterns and make predictions about future changes. Understanding trends not only helps us make informed decisions and develop predictive models but also allows us to anticipate changes and adapt strategies accordingly. Thus, a positive regression coefficient is an essential component of trend analysis and model evaluation. It provides valuable information on the long-term dynamics of the data as well as the direction in which the related variables are moving.

Several models available for analysing time series data, which can be used to detect patterns and trends in the data as well as to visualise them on graphs. However, it is important to consider both the external and internal characteristics of the data when choosing a model. With regard to this study, both logical autoregression and functional autoregression are methods that aim to detect key patterns in the data. However, their analytical approaches differ. Logical autoregression uses operations and rules to identify nonlinear dependencies, while functional autoregression adopts flexible approaches to capture complex functional changes. Effectively, the graphs created using these methods may look similar in terms of shape and direction, but a deeper analysis would reveal significant differences between the two. While logical graphs often show abrupt changes and anomalies, functional graphs are smoother and more natural. Therefore, it is important to choose the right model for the specific dataset being analysed. In particular, logical models may be more suitable for data with nonlinear dependencies, while functional models may be a better choice for analysing data with complex functional changes.

The quality of the models, as determined by the regression coefficient, is another important factor. When analysing the use of cluster methods and forecasting based on flat segments, pulsations and polynomials, it is crucial to consider the intersections of the logical and functional autoregressive trends. These intersections indicate areas of consistency between the methods, thus providing information about the reliability of the results.

Furthermore, to predict future trends, it is often necessary to analyse polynomial curves on graphs. The polynomial function plays a crucial role in predicting future trends and examining the significance of the data. Its shape and extrapolation can provide insights into the development of time series for the future. Therefore, special attention must be paid to polynomial functions for assessing future trends and forecasts based on time series. For instance, a second-degree polynomial function for trend analysis can be expressed as follows:

$$P\left(x\right) = a * x^2 + b * x + c \tag{14}$$

Where:

P(x) - value of the function at time x,

a, b, c - coefficients of the polynomial determining the shape of the curve,

x - time or period.

The polynomial function, defined as an equation that includes terms with powers greater than one, allows us to approximate the complex patterns of changes in data over time. Moreover, the analysis of the shape of the polynomial curve on the trend graphs of the logical and functional autoregressions allows for an accurate assessment of the prospects for development of the considered data in the future.

The analysis and evaluation of the resulting polynomial curve on trend charts represent the final step in predicting the relevance of the selected data in the future. The entire algorithm for identifying topics that are likely to be relevant in the future based on the functional and logical autoregression methods is presented in Figures 3 and 4.

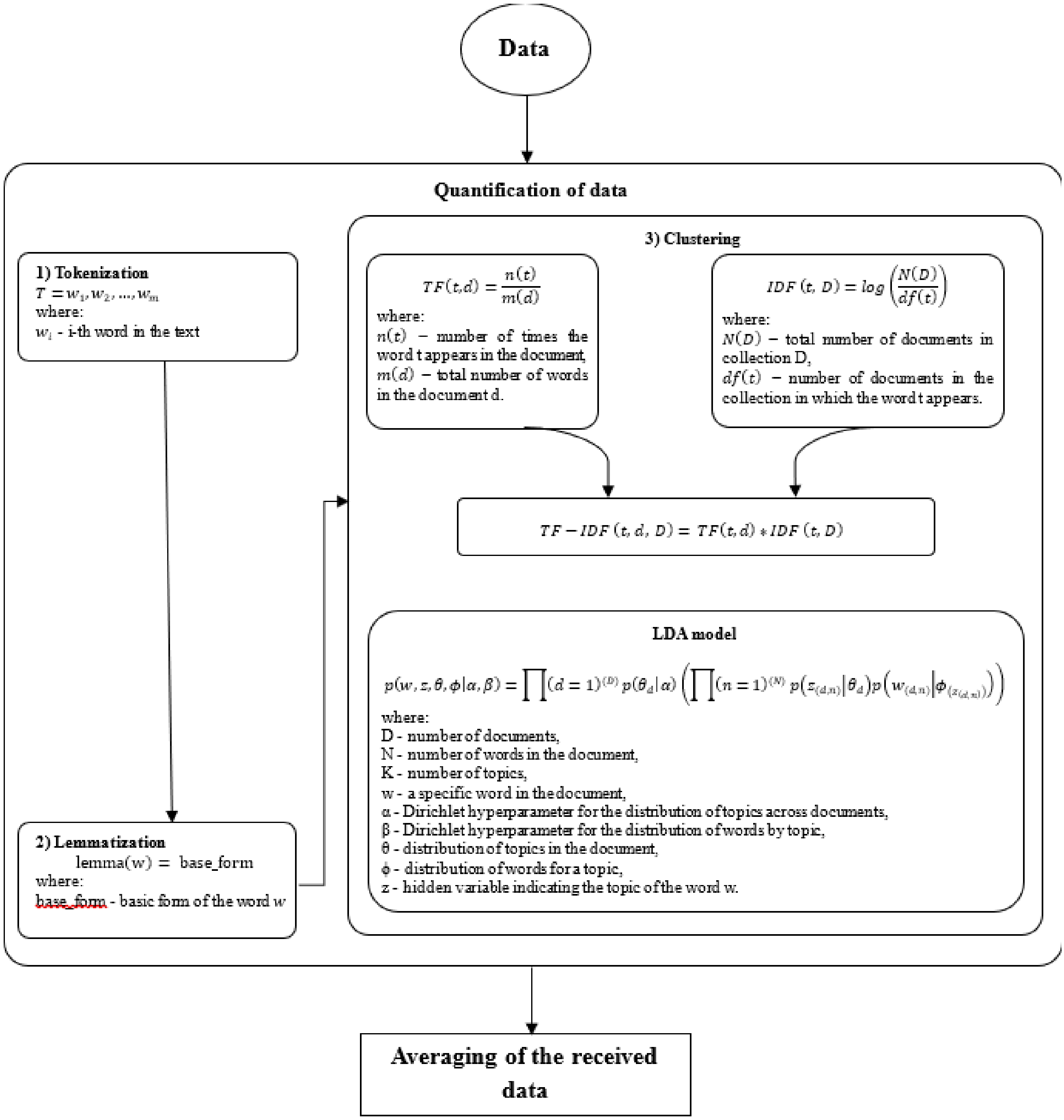

Data
Quantification of data
1) Tokenization
$T = w_1, w_2, \ldots, w_m$
where:
w_i - i-th word in the text
2) Lemmatization
lemma(w) = base_form
where:
base_form - basic form of the word w
3) Clustering
$TF(t,d) = \frac{n(t)}{m(d)}$
where:
$n(t)$ – number of times the word t appears in the document,
$m(d)$ – total number of words in the document d.
$IDF\ (t,\ D) = log\left(\frac{N(D)}{df(t)}\right)$
where:
$N(D)$ – total number of documents in collection D,
$df(t)$ – number of documents in the collection in which the word t appears.
$TF - IDF\ (t,\ d,\ D) = TF(t,d) * IDF\ (t,\ D)$
LDA model
$p(w, z, \theta, \phi | \alpha, \beta) = \prod(d = 1)^{(D)} p(\theta_d | \alpha) \left(\prod(n = 1)^{(N)} p(z_{(d,n)} | \theta_d) p\left(w_{(d,n)} | \phi_{(z_{(d,n)})}\right)\right)$
where:
D - number of documents,
N - number of words in the document,
K - number of topics,
w - a specific word in the document,
α - Dirichlet hyperparameter for the distribution of topics across documents,
β - Dirichlet hyperparameter for the distribution of words by topic,
θ - distribution of topics in the document,
ϕ - distribution of words for a topic,
z - hidden variable indicating the topic of the word w.
Averaging of the received data

Figure 3. First part of the algorithm for identifying relevant topics for the future

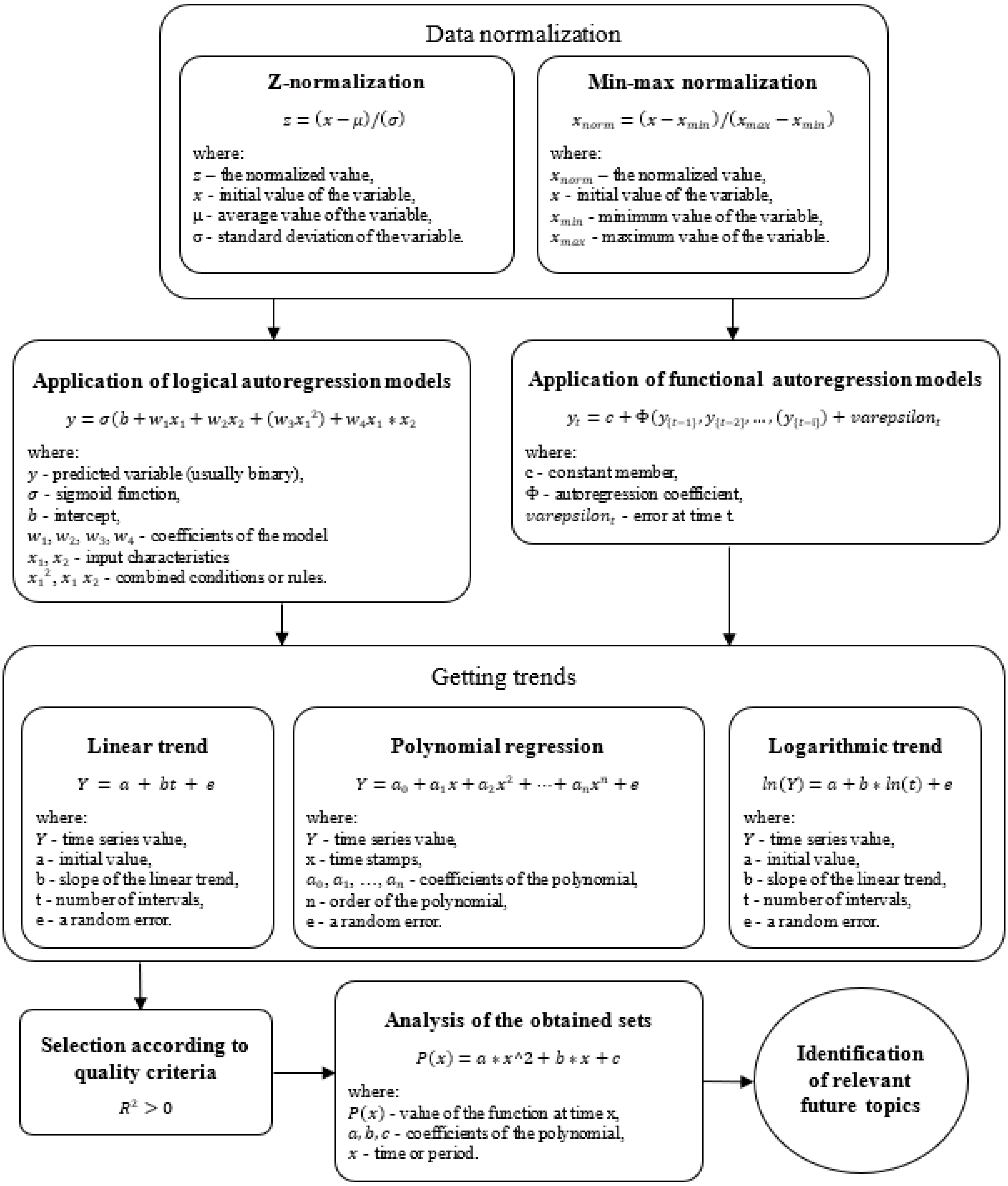

Data normalization
Z-normalization
$z = (x - \mu)/(\sigma)$
where:
z – the normalized value,
x - initial value of the variable,
μ - average value of the variable,
σ - standard deviation of the variable.
Min-max normalization
$x_{norm} = (x - x_{min})/(x_{max} - x_{min})$
where:
x_{norm} – the normalized value,
x - initial value of the variable,
x_{min} - minimum value of the variable,
x_{max} - maximum value of the variable.
Application of logical autoregression models
$y = \sigma(b + w_1x_1 + w_2x_2 + (w_3x_1^2) + w_4x_1 * x_2$
where:
y - predicted variable (usually binary),
σ - sigmoid function,
b - intercept,
w_1, w_2, w_3, w_4 - coefficients of the model
x_1, x_2 - input characteristics
$x_1{}^2, x_1\, x_2$ - combined conditions or rules.
Application of functional autoregression models
$y_t = c + \Phi(y_{\{t-1\}}, y_{\{t-2\}}, \dots, (y_{\{t-i\}}) + varepsilon_t$
where:
c - constant member,
Φ - autoregression coefficient,
$varepsilon_t$ - error at time t
Getting trends
Linear trend
$Y\ =\ a\ +\ bt\ +\ e$
where:
Y - time series value,
a - initial value,
b - slope of the linear trend,
t - number of intervals,
e - a random error.
Polynomial regression
$Y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n + e$
where:
Y - time series value,
x - time stamps,
$a_0, a_1, \dots, a_n$ - coefficients of the polynomial,
n - order of the polynomial,
e - a random error.
Logarithmic trend
$ln(Y) = a + b * ln(t) + e$
where:
Y - time series value,
a - initial value,
b - slope of the linear trend,
t - number of intervals,
e - a random error.
Selection according to quality criteria
$R^2 > 0$
Analysis of the obtained sets
$P(x) = a * x^\wedge 2 + b * x + c$
where:
$P(x)$ - value of the function at time x,
a, b, c - coefficients of the polynomial,
x - time or period.
Identification of relevant future topics

Figure 4. Second part of the algorithm for identifying relevant topics for the future

5. Discussion

As shown in the algorithm diagram in Figures 3 and 4, the process began with data quantification and normalisation. Both of these steps are important to ensure the high quality of the data before analysing them using autoregressive models. Notably, the main focus of this study was to apply autoregression methods to analyse long-term trends and predict future developments in the data through trend analysis. By building a trend using various methods, such as linear, logarithmic and polynomial regression, we were able to model the long-term dynamics of the data and predict its relevance in the future.

Modern information technologies provide researchers with powerful tools for extracting valuable information from large datasets. However, the use of these technologies requires rigorous data preprocessing, with normalisation being a crucial part of this process. Normalisation refers to the process of standardising variable values, which not only enhances the accuracy of forecasts but also leads to a better understanding of trends and patterns within the data.

The algorithm concludes with a polynomial function. By analysing the shape of the polynomial curve on trend graphs, we identified patterns of how the data changed over time, which can help make informed decisions based on forecasts. For instance, in the financial sector, this algorithm can help identify the most promising investment opportunities when prices fluctuate sharply. Furthermore, in the field of marketing and data analysis, studying the variability of topics could help identify current trends in consumer behaviour and user requests.

Thus, the combination of functional and logical autoregression models with polynomial regression trends provides a comprehensive approach for analysing and predicting time dependencies, thereby offering valuable information for strategic decision making and model development.

6. Conclusion

In this work, we developed an algorithm to identify topics that are likely to be relevant in the future. The first step in this process was data collection, followed by the calculation of average values and normalisation. Notably, these steps are crucial to ensure high-quality data for the application of autoregressive models. Data preprocessing significantly improves the accuracy and reliability of the analysis, providing the necessary foundation for subsequent modelling.

The main focus of this study was to implement autoregression methods to analyse long-term trends and predict future developments in the data. The proposed algorithm also incorporated an assessment of future forecasts, which is a novel approach to analysing emerging trends. By analysing graphical trends, we can explore, model and predict future data dynamics in greater precision. Furthermore, the regression coefficient obtained after constructing the trend line was chosen as the quality criterion in this study. This allowed us to analyse all the identified trends, their changes and opportunities for forecasting events and trends in the study area. The algorithm concluded with a polynomial function, which enabled the identification of relevant future topics for further research.

Overall, this algorithm serves as a powerful tool for analysing and forecasting future trends, helping to identify prospects for technological development based on historical data analysis.

References

Aizawa, A., 2003. An information-theoretic perspective of tf-idf measures. Inf. Process. Manag. 39(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Boban, I., Doko, A., Gotovac, S., 2020. Sentence retrieval using stemming and lemmatization with different length of the queries. Adv. Sci. Technol. Eng. Syst. 5(3). https://doi.org/10.25046/aj050345

Choi, J., Lee, S. W., 2020. Improving FastText with inverse document frequency of subwords. Pattern Recognit. Lett. 133. https://doi.org/10.1016/j.patrec.2020.03.003

Cover, T. M., Thomas, J. A., 2005. Elements of Information Theory. John Wiley and Sons, New York. https://doi.org/10.1002/047174882X

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., Jain, A., 2024. Structured information extraction from scientific text with large language models. Nat. Commun. 15(1). https://doi.org/10.1038/S41467-024-45563-X

Dey, R. K., Das, A. K., 2023. Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis. Multim. Tools Appl. 82(21). https://doi.org/10.1007/s11042-023-14653-1

Di, Y., Zhang, Y., Zhang, L., Tao, T., Lu, H., 2017. MdFDIA: A mass defect based four-plex data-independent acquisition strategy for proteome quantification. Anal. Chem. 89(19), 10248–10255. https://doi.org/10.1021/acs.analchem.7b01635

Friedman, R., 2023. Tokenization in the theory of knowledge. Encycl. 3(1). https://doi.org/10.3390/encyclopedia3010024

Gandhi, A. B., Joshi, J. B., Kulkarni, A. A., Jayaraman, V. K., Kulkarni, B. D., 2008. SVR-based prediction of point gas hold-up for bubble column reactor through recurrence quantification analysis of LDA time-series. Int. J. Multiph. Flow 34(12), 1099–1107. https://doi.org/10.1016/j.ijmultiphaseflow.2008.07.001

Huang, G. bin, Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man Cybern. B Cybern. 42(2), 513–529. https://doi.org/10.1109/TSMCB.2011.2168604

Huang, Q., Zhang, H., Chen, J., He, M., 2017. Quantile regression models and their applications: A review. J. Biom. Biostat. 8(3). https://doi.org/10.4172/2155-6180.1000354

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimed. Tools Appl. 78(11). https://doi.org/10.1007/s11042-018-6894-4

Kadiev, I. P., Kadiev, P. A., 2016. Homogeneous register environments with a programmable structure. Bull. Dagestan State Tech. Univ. Tech. Sci. 35(4), 108–112. https://doi.org/10.21822/2073-6185-2014-35-4-108-112

Mahmoud, H. A. H., Hafez, A. M., Alabdulkreem, E., 2023. Language-independent text tokenization using unsupervised deep learning. Intell. Autom. Soft Comput. 35(1). https://doi.org/10.32604/iasc.2023.026235

Mestre, G., Portela, J., Rice, G., Muñoz San Roque, A., Alonso, E., 2021. Functional time series model identification and diagnosis by means of auto- and partial autocorrelation analysis. Comput. Stat. Data Anal. 155, 107108. https://doi.org/10.1016/J.CSDA.2020.107108

Minogue, C. E., Hebert, A. S., Rensvold, J. W., Westphall, M. S., Pagliarini, D. J., Coon, J. J., 2015. Multiplexed quantification for data-independent acquisition. Anal. Chem. 87(5), 2570–2575. https://doi.org/10.1021/AC503593D

Ozturkmenoglu, O., Alpkocak, A., 2012. Comparison of different lemmatization approaches for information retrieval on Turkish text collection. Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications. https://doi.org/10.1109/INISTA.2012.6246934

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Puchkov, E. V., Belyavsky, G. I., 2018. The use of local trends for the pre-preparation of time series in forecasting tasks. Int. J. Softw. Prod. Syst. 29, 751–756. https://doi.org/10.15827/0236-235X.124.751-756

Savzikhanova, S. A., 2023. Big data is a winning innovation for predicting future trends. UEPS: Management, Economics, Politics, Sociology, 69–75. https://doi.org/10.24412/2412-2025-2023-2-69-76

Shantal, M., Othman, Z., Bakar, A. A., 2023. A novel approach for data feature weighting using correlation coefficients and min–max normalization. Symmetry, 15(12). https://doi.org/10.3390/sym15122185

Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 97. https://doi.org/10.1016/j.asoc.2019.105524

Toporkov, O., Agerri, R., 2024. On the role of morphological information for contextual lemmatization. Comput. Linguist. 50(1). https://doi.org/10.1162/coli_a_00497

Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K. A., Ceder, G., Jain, A., 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns 3(4). https://doi.org/10.1016/J.PATTER.2022.100488

Zhang, B., Kä, L., Zubarev, R. A., 2016. DeMix-Q: Quantification-centered data processing workflow. Mol. Cell. Proteom. 15(4), 1467–1478. https://doi.org/10.1074/MCP.O115.055475

Zhang, W., Wang, Q., Kong, X., Xiong, J., Ni, S., Cao, D., Niu, B., Chen, M., Li, Y., Zhang, R., Wang, Y., Zhang, L., Li, X., Xiong, Z., Shi, Q., Huang, Z., Fu, Z., Zheng, M., 2024. Fine-tuning large language models for chemical text mining. Chem. Sci. 15(27), 10600–10611. https://doi.org/10.1039/d4sc00924j

Zhang, Z., Lei, Y., Xu, J., Mao, X., Chang, X., 2019. TFIDF-FL: Localizing faults using term frequency-inverse document frequency and deep learning. IEICE Trans. Inf. Syst. E102D(9). https://doi.org/10.1587/transinf.2018EDL8237

Список источников

Aizawa, A., 2003. An information-theoretic perspective of tf-idf measures. Information Processing and Management, 39(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Boban, I., Doko, A., & Gotovac, S., 2020. Sentence retrieval using Stemming and Lemmatization with different length of the queries. Advances in Science, Technology and Engineering Systems, 5(3). https://doi.org/10.25046/aj050345

Choi, J., & Lee, S. W., 2020. Improving FastText with inverse document frequency of subwords. Pattern Recognition Letters, 133. https://doi.org/10.1016/j.patrec.2020.03.003

Cover, T. M., & Thomas, J. A., 2005. Elements of Information Theory. In Elements of Information Theory. John Wiley and Sons. https://doi.org/10.1002/047174882X

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A., 2024. Structured information extraction from scientific text with large language models. Nature Communications, 15(1). https://doi.org/10.1038/S41467-024-45563-X

Dey, R. K., & Das, A. K., 2023. Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis. Multimedia Tools and Applications, 82(21). https://doi.org/10.1007/s11042-023-14653-1

Di, Y., Zhang, Y., Zhang, L., Tao, T., & Lu, H., 2017. MdFDIA: A Mass Defect Based Four-Plex Data-Independent Acquisition Strategy for Proteome Quantification. Analytical Chemistry, 89(19), 10248–10255. https://doi.org/10.1021/acs.analchem.7b01635

Friedman, R., 2023. Tokenization in the Theory of Knowledge. Encyclopedia, 3(1). https://doi.org/10.3390/encyclopedia3010024

Gandhi, A. B., Joshi, J. B., Kulkarni, A. A., Jayaraman, V. K., & Kulkarni, B. D., 2008. SVR-based prediction of point gas hold-up for bubble column reactor through recurrence quantification analysis of LDA time-series. International Journal of Multiphase Flow, 34(12), 1099–1107. https://doi.org/10.1016/j.ijmultiphaseflow.2008.07.001

Huang, G. bin, Zhou, H., Ding, X., & Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 42(2), 513–529. https://doi.org/10.1109/TSMCB.2011.2168604

Huang, Q., Zhang, H., Chen, J., & He, M., 2017. Quantile Regression Models and Their Applications: A Review. Journal of Biometrics & Biostatistics, 08(03). https://doi.org/10.4172/2155-6180.1000354

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11). https://doi.org/10.1007/s11042-018-6894-4

Mahmoud, H. A. H., Hafez, A. M., & Alabdulkreem, E., 2023. Language-Independent Text Tokenization Using Unsupervised Deep Learning. Intelligent Automation and Soft Computing, 35(1). https://doi.org/10.32604/iasc.2023.026235

Mestre, G., Portela, J., Rice, G., Muñoz San Roque, A., & Alonso, E., 2021. Functional time series model identification and diagnosis by means of auto- and partial autocorrelation analysis. Computational Statistics & Data Analysis, 155, 107108. https://doi.org/10.1016/J.CSDA.2020.107108

Minogue, C. E., Hebert, A. S., Rensvold, J. W., Westphall, M. S., Pagliarini, D. J., & Coon, J. J., 2015. Multiplexed quantification for data-independent acquisition. Analytical Chemistry, 87(5), 2570–2575. https://doi.org/10.1021/AC503593D

Ozturkmenoglu, O., & Alpkocak, A., 2012. Comparison of different lemmatization approaches for information retrieval on Turkish text collection. International Symposium on Innovations in Intelligent SysTems and Applications. https://doi.org/10.1109/INISTA.2012.6246934

Peng, H., Long, F., & Ding, C., 2005. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226–1238.

https://doi.org/10.1109/TPAMI.2005.159

Shantal, M., Othman, Z., & Bakar, A. A., 2023. A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. Symmetry, 15(12). https://doi.org/10.3390/sym15122185

Singh, D., & Singh, B., 2020. Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97. https://doi.org/10.1016/j.asoc.2019.105524

Toporkov, O., & Agerri, R., 2023. On the Role of Morphological Information for Contextual Lemmatization. Computational Linguistics, 50(1). https://doi.org/10.1162/coli_a_00497

Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K. A., Ceder, G., & Jain, A., 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns (New York, N.Y.), 3(4). https://doi.org/10.1016/J.PATTER.2022.100488

Zhang, B., Kä, L., & Zubarev, R. A., 2016. DeMix-Q: Quantification-Centered Data Processing Workflow. Molecular & Cellular Proteomics : MCP, 15(4), 1467–1478. https://doi.org/10.1074/MCP.O115.055475

Zhang, W., Wang, Q., Kong, X., Xiong, J., Ni, S., Cao, D., Niu, B., Chen, M., Li, Y., Zhang, R., Wang, Y., Zhang, L., Li, X., Xiong, Z., Shi, Q., Huang, Z., Fu, Z., & Zheng, M., 2024. Fine-tuning large language models for chemical text mining. Chemical Science, 15(27), 10600–10611. https://doi.org/10.1039/d4sc00924j

Zhang, Z., Lei, Y., Xu, J., Mao, X., & Chang, X., 2019. TFIDF-FL: Localizing faults using term frequency-inverse document frequency and deep learning. IEICE Transactions on Information and Systems, E102D(9). https://doi.org/10.1587/transinf.2018EDL8237

Кадиев, И. П., & Кадиев, П. А., 2016. Однородные регистровые среды с программируемой структурой. Вестник Дагестанского Государственного Технического Университета. Технические Науки, 35(4), 108–112. https://doi.org/10.21822/2073-6185-2014-35-4-108-112

Пучков, Е. В., Puchkov, E. v., Белявский, Г. И., & Belyavsky, G. I., 2018. Применение локальных трендов для предподготовки временных рядов в задачах прогнозирования. Международный Журнал Программные Продукты и Системы, 29, 751–756. https://doi.org/10.15827/0236-235X.124.751-756

Савзиханова, С.А., 2023, Big Data – выигрышная инновация для прогнозирования будущих тенденций. УЭПС: управление, экономика, политика, социология, 69–75. https://doi.org/10.24412/2412-2025-2023-2-69-76

The article was submitted 23.07.2024, approved after reviewing 6.08.2024, accepted for publication 20.08.2024.

Статья поступила в редакцию 23.07.2024, одобрена после рецензирования 6.08.2024, принята к публикации 20.08.2024.

About the authors:

1. Aleksandr Borzov, chancellor, St. Petersburg Restoration and Construction Institute, St. Petersburg, Russia. https://orcid.org/0009-0005-2869-8812, priem@spbrsi.ru

2. Viktor Sorokin, laboratory assistant, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia. https://orcid.org/0009-0007-5061-9636, vitya.soroka.02@mail.ru

3. Vitaly Mityazov, laboratory assistant, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia. https://orcid.org/0009-0004-6573-554X, vitalikmit@mail.ru

4. Daria Shimanova, researcher, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia. shimanova_da@spbstu.ru

Информация об авторах:

1. Александр Борзов, ректор, Санкт-Петербургский реставрационно-строительный институт, Санкт-Петербург, Россия. https://orcid.org/0009-0005-2869-8812, priem@spbrsi.ru

2. Виктор Сорокин, лаборант, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия. https://orcid.org/0009-0007-5061-9636, vitya.soroka.02@mail.ru

3. Виталий Митязов, лаборант, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия. https://orcid.org/0009-0004-6573-554X, vitalikmit@mail.ru

4. Дарья Шиманова, ассистент, Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия. shimanova_da@spbstu.ru