ISSN: 3071-186X Volume 01, Issue 05, June, 2026

Scienta Press Journals

American Innovator
Journal of Emerging Technologies and Research

Methods Of Intellectual Analysis Of Textual Data

I.Sh. Ochildiyev.

Institute of Information Communication Technologies and Military signals

Abstract

This article examines modern methods of intellectual analysis of textual data, their theoretical foundations, and practical applications. Particular attention is given to text processing, data classification, clustering, sentiment analysis, topic modeling, and knowledge extraction using artificial intelligence and machine learning algorithms. The study highlights the role of Natural Language Processing (NLP) technologies in handling the rapidly growing volume of textual information. Various approaches to extracting meaningful insights from unstructured text are analyzed, and their effectiveness in supporting decision-making processes is evaluated. The results demonstrate the potential of intelligent text analysis methods for improving information management, knowledge discovery, and data-driven decision-making across different domains.

Keywords: Textual Data, Intelligent Analysis, Artificial Intelligence, Machine Learning, Natural Language Processing, Text Classification, Clustering, Sentiment Analysis, Knowledge Extraction, Data Mining.

Аннотация: В данной статье рассматриваются современные методы интеллектуального анализа текстовых данных, их теоретические основы и практическое применение. Особое внимание уделяется обработке текстов, классификации данных, кластеризации, анализу тональности, тематическому моделированию, а также извлечению знаний с использованием алгоритмов искусственного интеллекта и машинного обучения. В исследовании подчеркивается роль технологий обработки естественного языка (NLP) в работе с быстро растущими объемами

текстовой информации. Проанализированы различные подходы к извлечению содержательных знаний из неструктурированных текстов и оценена их эффективность в поддержке процессов принятия решений. Результаты демонстрируют высокий потенциал методов интеллектуального анализа текстов для совершенствования управления информацией, выявления новых знаний и принятия решений на основе данных в различных сферах деятельности.

Ключевые слова: текстовые данные, интеллектуальный анализ, искусственный интеллект, машинное обучение, обработка естественного языка, классификация текстов, кластеризация, анализ тональности, извлечение знаний, интеллектуальный анализ данных.

Annotatsiya: Ushbu maqolada matnli ma’lumotlarni intellektual tahlil qilishning zamonaviy usullari, ularning nazariy asoslari va amaliy qo‘llanilishi ko‘rib chiqilgan. Matnlarni qayta ishlash, ma’lumotlarni tasniflash, klasterlash, hissiyotlarni tahlil qilish, mavzularni modellashtirish hamda sun’iy intellekt va mashinaviy o‘qitish algoritmlari yordamida bilimlarni ajratib olish masalalariga alohida e’tibor qaratilgan. Tadqiqotda tabiiy tilni qayta ishlash (NLP) texnologiyalarining tez sur’atlarda ortib borayotgan matnli axborot hajmini qayta ishlashdagi o‘rni yoritilgan. Tuzilmagan matnlardan mazmunli bilimlarni ajratib olishning turli yondashuvlari tahlil qilinib, ularning qaror qabul qilish jarayonlarini qo‘llab-quvvatlashdagi samaradorligi baholangan. Natijalar intellektual matn tahlili usullarining axborotni boshqarish, yangi bilimlarni aniqlash va ma’lumotlarga asoslangan qarorlar qabul qilishni takomillashtirishdagi katta salohiyatini namoyon etadi.

Kalit so‘zlar: Matnli ma’lumotlar, intellektual tahlil, sun’iy intellekt, mashinaviy o‘qitish, tabiiy tilni qayta ishlash, matnlarni tasniflash, klasterlash, hissiyotlarni tahlil qilish, bilimlarni ajratib olish, ma’lumotlarni qazib olish.

The rapid growth of digital technologies has led to an unprecedented increase in the volume of textual data generated through social networks, websites, electronic documents, online communications, scientific publications, and information systems. As a result, the effective processing and analysis of textual information have become important challenges for researchers, organizations, and decision-makers. Traditional methods of data analysis are often insufficient for handling large-scale unstructured text collections, creating a need for advanced intelligent techniques capable of extracting meaningful knowledge from textual data.

Methods of intellectual analysis of textual data combine approaches from Artificial Intelligence (AI), Machine Learning (ML), Data Mining, and

Natural Language Processing (NLP) to identify patterns, relationships, and hidden information within text documents. These methods enable the automated processing of vast amounts of textual information, reducing the time and effort required for manual analysis. Furthermore, intelligent text analysis techniques support various tasks, including text classification, clustering, sentiment analysis, topic detection, information retrieval, and knowledge extraction.

In recent years, significant advances in machine learning algorithms and deep learning models have substantially improved the accuracy and efficiency of text analysis systems. Modern NLP technologies allow computers to understand, interpret, and generate human language with a high degree of precision. The emergence of large language models and transformer-based architectures has further expanded the capabilities of intelligent text analytics, enabling more sophisticated semantic analysis and contextual understanding.

The importance of intellectual text analysis continues to grow across numerous domains, including cybersecurity, healthcare, education, finance, e-commerce, and public administration. Organizations increasingly rely on these methods to discover valuable insights, monitor public opinion, detect emerging trends, and support strategic decision-making processes. Therefore, the study of intelligent methods for textual data analysis remains a relevant and rapidly developing research area.

The technology of Data Mining is an interdisciplinary field of research that emerged and has been developed based on theories of probability and applied statistics, information theory, pattern recognition, artificial intelligence, database theory, data warehouses, high-performance computing, data visualization, machine learning, and others [1].

The concept of Data Mining, which appeared in 1978, gained significant popularity in its modern interpretation from the first half of the 1990s. The term “Data Mining” is often translated as data extraction, information extraction, data digging, intelligent data analysis, tools for finding patterns among information elements, knowledge extraction, pattern analysis, and other less common synonyms [2].

Since Data Mining is an interdisciplinary field, several definitions of this term have been proposed, including the following: Data Mining is the process of discovering previously unknown, non-trivial, practically useful, and interpretable knowledge in raw data, which is necessary for decision-making in various fields of human activity. It is also defined as “the process of extracting implicit and unstructured information from data and presenting it in a form suitable for use” [3].

It should be noted that the term “Data mining” is somewhat misleading, since the goal is not the extraction (mining) of data itself, but rather the extraction of patterns and knowledge from large volumes of data. To

provide a clear definition of data mining, it is necessary to distinguish between the following concepts, which, although used as synonyms, have fundamental differences.

The term “data” originates from the word data meaning fact, while information refers to explanation or presentation, i.e., facts or messages. According to [4], data is a collection of operational and objective facts describing objects, events, phenomena, processes, etc. Information is a selected, organized, and contextually processed part of data endowed with a system of relationships between data and a specific meaning (data about data, or data plus metadata). Knowledge is a complex network hierarchy of information elements with identified dependencies and/or significant relationships between facts, events, phenomena, and processes, etc. [5].

Therefore, the essence and purpose of Data Mining technology can be described as follows:

It is a technology designed to search large volumes of data for non-obvious, objective, and practically useful patterns.

Non-obvious means that the discovered patterns cannot be identified using standard information processing methods or expert analysis.

Objective means that the discovered patterns fully correspond to reality, unlike expert opinion, which is always subjective.

Practically useful means that the obtained conclusions have concrete significance and can be applied in practice.

Knowledge is a set of information that forms a coherent description corresponding to a certain level of awareness about a subject, problem, or domain.

In this section, the author has provided definitions of the main terms necessary for understanding the subject area of the dissertation research. The next section presents an analytical review of the main tasks and methods of Data Mining technology.

Data Mining tasks are divided into descriptive and predictive tasks.

Predictive tasks involve predicting unknown data values using known values. These include forecasting, classification, regression, etc. Descriptive tasks identify patterns or relationships in data and explore data properties [6]. These include clustering, association rules, sequence discovery, etc.

Below, the most common Data Mining tasks are discussed, with special attention given to classification and clustering, since they are used in this research for text analysis.

Classification is the derivation of a model that determines the class of an object based on its attributes [2]. A set of objects is defined as a training set, where each object is represented by a feature vector along with its class label.

ISSN: 3071-186X Volume 01, Issue 05, June, 2026

By analyzing the relationship between object attributes and classes in the training set, a classification model can be built. This task is an example of supervised learning, since classes are predefined before learning from target data.

As a result of classification, features that characterize groups of objects (classes) are identified, and based on these features, a new object can be assigned to a specific class.

There are many classification methods, the most well-known being: k-nearest neighbors ($k-NN$), decision trees, genetic algorithms, support vector machines (SVM), and neural networks [3].

Below, classifiers relevant to the methods and algorithms developed in this research are described in detail.

The Support Vector Machine (SVM) uses a supervised learning procedure for classification and regression tasks [6]. It works by finding the best hyperplane that separates a dataset, where a hyperplane is a subspace that divides data into different classes.

SVM is popular due to its high accuracy, computational efficiency, and ability to solve nonlinear problems using a kernel function, which maps data from one space into another. This function allows the method to operate in a higher-dimensional feature space.

Different types of kernel functions are used in SVM, including linear, nonlinear, polynomial (poly), radial basis function (RBF), and sigmoid kernels. These functions allow SVM to solve a wide range of problems, including those where classes are not linearly separable.

Support vectors are the points closest to the hyperplane. A hyperplane is an $(n-1)$-dimensional subspace used to divide the space into multiple regions.

The SVM method aims to find the optimal decision boundary between two classes with the highest generalization ability [7]. Optimality is determined by maximizing the margin, which is the distance between data points and the decision boundary. The margin width is crucial for SVM, as a larger margin leads to better generalization.

Figure 1 shows two classes separated by a hyperplane with a maximum margin.

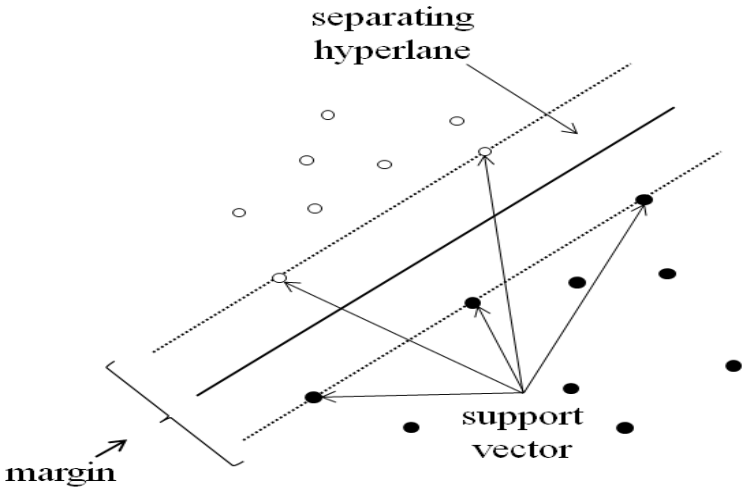

Figure 1 – Two (linearly separable) classes separated by a hyperplane with maximum margin relative to support vectors

The optimization problem for finding the separating hyperplane involves minimizing ||w|| under the constraints
$y_i = (w * x_i - b) \geq 1$ for all training examples, where w is the weight vector of the hyperplane, b is the bias term, and y_i is the class label of the training example.
The vector w is obtained by solving an optimization problem using standard quadratic programming methods, such as sequential minimal optimization or stochastic gradient descent. It can be expressed in the following form using Lagrange multipliers α\alphaα:

$$\arg\min_{w,b}\max_{a\geq 0}\left\{\frac{1}{2}||w||^2 - \sum_i^n \alpha_i[y_i(\langle w, x_i\rangle - b) - 1\}]\right\} \qquad (1)$$

One of the classifiers selected for conducting computational experiments in this study is the Support Vector Machine (SVM), as it demonstrated better performance in text classification tasks compared to other methods. Artificial Neural Networks (ANNs) are a type of machine learning algorithm inspired by the structure and function of the human brain. They consist of interconnected nodes (neurons) organized into layers. The entire logic of ANN operation is embedded in the connections between neurons [4].
Artificial neurons are the fundamental building blocks of ANNs. Each neuron receives input signals from other neurons or external sources, processes them using an activation function, and produces an output signal that is transmitted to connected neurons. Activation functions introduce non-linearity into ANNs, enabling them to model complex relationships between input and output data. Common activation functions include sigmoid, tanh, and ReLU [2].

Weights and biases are numerical values associated with connections between neurons. Weights determine the strength of the signal transmitted from one neuron to another, while biases adjust the overall activation level of a neuron.

ANNs are organized into layers, where each layer consists of a group of interconnected neurons. A typical ANN architecture includes an input layer, one or more hidden layers, and an output layer. The output of each neuron is determined by a weighted sum of its inputs, followed by the application of an activation function. This process can be mathematically represented as follows:

$$z = \sum_{i=1}^{n}(w_i * x_i) + b, \quad (2)$$

where z is the weighted sum of inputs and bias, w_i are the weights associated with the input features, x_i are the input values, and $\mathbf{b}$ is the bias term.

Then, the output of the neuron is passed through an activation function such as the sigmoid function, tanh, or ReLU in order to introduce non-linearity into the network.

The output of the neuron **a** is defined by the following expression:

$$a = f(z) \quad (3)$$

Artificial Neural Networks (ANNs) are trained using a set of labeled examples, where each example consists of an input sample and the corresponding desired output. The training process includes forward propagation, error computation, backpropagation, and optimization using algorithms such as gradient descent.

Artificial Neural Networks have found wide application in various fields, including image recognition, predictive modeling, and natural language processing (NLP), such as machine translation, sentiment analysis, and text summarization.

The neural network classifier will be used in computational experiments to evaluate the effectiveness of working with vectors obtained through the developed methods (see Section 4.1).

The k-Nearest Neighbors (k-NN) algorithm is a non-parametric supervised learning method used for classification and regression tasks. It is based on the assumption that similar objects are located close to each other in the feature space [8].

In the case of classification, the class of a new data point is determined by the majority vote of its k nearest neighbors. Mathematically, this can be represented as follows:

$$\hat{y} = \text{majority vote}\left(\{y_i\}_{i=1}^{k}\right), \quad (4)$$

where $(\hat{y})$ is the predicted class of a new data point, and (y_i) are the classes of the k nearest neighbors.

In the case of regression, the value of a new data point is estimated based on the average of its k nearest neighbors. Mathematically, this can be represented as follows:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i, \quad (5)$$

where $(\hat{y})$ is the calculated value for a new data point, and (y_i) are the values of the k nearest neighbors.

The choice of the value of k is crucial, as it significantly affects the performance of the algorithm. Various distance metrics can be used to measure similarity between data points, such as Euclidean distance, Manhattan distance, Minkowski distance, and Hamming distance [3].

The $k - N$ algorithm requires significant computational resources for large training datasets, as it computes distances from the test vector to all stored examples. However, this does not prevent its use in experiments for evaluating the effectiveness of the proposed methods.

Clustering is similar to classification, except that the groups are not predefined but are instead determined by analyzing the structure and behavior of the data (unsupervised learning). The result of clustering is the partitioning of objects into groups. There are many clustering methods, including K-means, DBSCAN, hierarchical clustering, spectral clustering, and others [8].

Below is a detailed explanation of the K-means algorithm and its working mechanism, as it will be used to construct and evaluate the effectiveness of the proposed text analysis methods in this study.

The K-means clustering algorithm is a popular unsupervised machine learning method used to group unlabeled datasets into different clusters [9]. It is a centroid-based algorithm that aims to divide n observations into $k\ (\leq n)$ sets (S = {S_1, S_2, ..., S_k}) in order to minimize the sum of squared within-cluster distances (i.e., variance). Formally, the objective can be represented by the following expression:

$$\operatorname{argmin}_S \sum_{i=1}^{k} \sum_{x \in s_i} ||x - \mu_i||^2, \quad (6)$$

where S is a set of k clusters, x is a data point, and $(\mu_i)$is the mean value of the data points in the i cluster.

The algorithm consists of the following steps:

1. Random initialization of k cluster centroids.
2. Assigning each data point to the nearest centroid.
3. Recomputing the centroids as the mean of the data points assigned to each cluster.
4. Repeating steps 2 and 3 until convergence.

The K-means algorithm is sensitive to the initial placement of centroids and may converge to a local minimum. To overcome this drawback, the $\mathrm{K-means}$ algorithm can be used, which initializes centroids in a way that generally produces better results than random initialization [10].

The $\mathrm{K-means}$ algorithm has various applications, including image segmentation, document clustering, and market segmentation.

Another variant of $\mathrm{K-means}$ is the Spherical K-means algorithm, which is particularly effective for processing sparse and high-dimensional data such as document vectors. This algorithm differs in that it uses cosine similarity as the distance measure between vectors and is therefore used to implement the proposed vectorization method by applying it during the stage of concept extraction from text.

The reason for choosing spherical $\mathrm{K-means}$ is that cosine similarity is more effective compared to the standard K-means approach. In addition, the objective function in spherical $\mathrm{K-means}$ is better suited to the idea of a centroid as a concept representation. As for evaluating the effectiveness of the developed text vectorization method in the text clustering task, the standard $\mathrm{K-means}$ algorithm is used due to its flexibility and scalability, which makes this choice more suitable for different types of vectors.

For the classification task, preliminary experiments were conducted using classifiers such as Artificial Neural Networks (ANN), Random Forest (RF), and Support Vector Machines (SVM) on vectors generated by the proposed methods. SVM showed the best results and is one of the widely accepted approaches for evaluating the discriminative ability of feature vectors. Therefore, it will be used in all experiments to evaluate the performance of the proposed methods compared to other approaches.

This section described classification and clustering tasks, as well as analyzed methods for solving them, which will be applied in this study for building the proposed methods and in computational experiments for evaluating their effectiveness.

Conclusion

The intellectual analysis of textual data has become an essential component of modern information processing systems due to the rapid growth of unstructured textual information. Advanced methods based on artificial intelligence, machine learning, data mining, and natural language processing enable the extraction of valuable knowledge, patterns, and insights from large text collections. These techniques significantly improve the efficiency and accuracy of tasks such as text classification, clustering, sentiment analysis, topic detection, and information retrieval. Recent developments in deep learning and transformer-based models have

further enhanced the capabilities of intelligent text analysis by providing a deeper understanding of linguistic and semantic relationships within textual data. The application of these methods spans numerous domains, including cybersecurity, healthcare, education, finance, business intelligence, and public administration. Despite existing challenges related to data quality, language ambiguity, and computational complexity, continuous advancements in artificial intelligence are expanding the effectiveness of text analytics solutions. Therefore, the development and implementation of intelligent textual data analysis methods remain a promising research direction that contributes to improved decision-making, knowledge discovery, and information management in the digital era.

References:

[1] Speech and Language Processing / Daniel Jurafsky, James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd ed. Stanford University, 2024.
[2] Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
[3] Data Mining Concepts and Techniques / Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann, 2011.
[4] Mining of Massive Datasets / Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. 3rd ed. Cambridge University Press, 2020.
[5] Foundations of Statistical Natural Language Processing / Christopher D. Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
[6] Natural Language Processing with Python / Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'Reilly Media, 2009.
[7] Deep Learning / Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep Learning. MIT Press, 2016.
[8] Attention Is All You Need / Ashish Vaswani et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems (NeurIPS), 2017.
[9] BERT Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT, 2019.

ISSN: 3071-186X Volume 01, Issue 05, June, 2026

[10] Text Mining Applications and Theory / Michael W. Berry, Jacob Kogan. Text Mining: Applications and Theory. Wiley, 2010.