International Journal of Technology v(i) pp-pp (YYYY)
Received Month Year / Revised Month Year / Accepted Month Year

International Journal of Technology

http://ijtech.eng.ui.ac.id

Functional Modeling of Distributions of Substantive-Content Message Properties in the Information Background

Evgenii Konnikov[1], Dmitriy Rodionov[1], Darya Kryzhko[*1]

[1] *Peter the Great St. Petersburg Polytechnic University, 29 Politechnicheskaya Ulitsa, St. Petersburg, 195251 Russia*

Abstract. This paper presents a novel methodology for modeling the distribution of substantive-content message properties in the information background. This study develops a toolkit to analyze and predict information dynamics by identifying key themes, evaluating their importance, and understanding their connections.. The proposed approach is based on the concept of multimodality, where properties are characterized by peaks of varying intensity and frequency. Intensity, reflecting the depth and saturation of the information signal, is modeled using the Gamma distribution, while frequency, reflecting the number of occurrences, is represented by the multivariate normal distribution. A genetic algorithm is employed to identify the optimal parameters for these distributions. The methodology offers a more comprehensive understanding of information dynamics by considering both intensity and frequency, and effectively handles complex interdependencies between properties. It can be applied to various domains, including social media analysis, political communication, and marketing, providing valuable insights for decision-making.

Keywords: Information Background Analysis; Intensity of Presence; Substantive-Content Message; Symbolic Data Constructs; Thematic Analysis

1. Introduction

The rapid changes in socio-economic relations, characteristic of the past few decades, are shaping new modern trends, rules, and laws governing the existence of enterprises. Today, in the period of the formation of a post-industrial economy, the significance of material factors for an enterprise corresponds to a smaller share of value in terms of the resource structure of activities than in the previous stages of economic development. Meanwhile, the weight of intangible elements, such as information, has a steadily positive trend towards growth (Zaytsev et al.,2019; Hitt et al.,2019; Damanpour, 2018; Berawi, M.A. at al., 2022; Eremina, I. et al., 2022; Nugraha, E. et al., 2022).

The stability of these trends is explained by the formation of an empirically and scientifically confirmed understanding of the comprehensive essence of information. Information is an intangible resource that breaks barriers and drives globalization. (Rodionov et al., 2021; Mueller et al.,2019; Bharadwaj et al., 2000).

.

*Corresponding author's email: darya.kryz@yandex.ru, Tel.: +00-00-000000; Fax: +00-00-000000
doi: 10.14716/ijtech.v0i0.0000

Information, presented as applied knowledge, determines the effectiveness of other classical factors of production (Drucker, 1990). Information has a huge impact on classical production resources such as labor, capital, and land (Anisiforov et al., 2017). It plays a role in resource management, production processes, and decision-making, which can lead to increased efficiency in the use of these resources and improved productivity (Öberg, 2023; Qiu et al.,2022; Rodionov et al., 2022).

Practical information obtained through the analysis of large data sets contributes to an increase in the quality of decision-making, both for individual actors within an enterprise and for the enterprise as a whole (Kushwaha et al., 2021). With the rapid increase in computing power available to analysts, the use of big data analytics for efficiency in small areas of everyday management is becoming a competitive necessity for organizations (Anisiforov et al., 2017). The costs of obtaining actionable information through the volume, variety, and veracity of data, enabling individuals or organizations to make informed, qualitatively balanced decisions underpinned by the fact of reducing the risk of uncertain outcomes, correlate with the income generated by the socio-economic benefits of such an approach (Rodionov et al., 2022; Cawsey et al., 2016). Therefore, studying the mechanism of information background formation among a group of subjects and developing a methodology for modeling the properties of the substantive-content message that arises in such an environment is becoming a pressing issue in contemporary research.

Karl Deutsch, an American political scientist and sociologist who studied the specifics of communication in the context of the state of political systems and management processes, posited the thesis that information has various properties that determine its impact on the system consuming it. Karl Deutsch was the first to begin to separately highlight what he called the "tone" or "temperature" of information, characterizing its emotional coloring. Deutsch also emphasized the consequences of the impact of information on society, in particular its ability to influence the status quo, to mobilize actions or suppress activity, and to affect the distribution of resources and power (Deutsch, 1951). As we can see, Deutsch assumed that information is not solely a factual or neutral set of data, but rather a substance carrying a significant burden in terms of impact and changes in social dynamics.

Fred Dretske, a pioneer in the philosophy of cognitive science, in turn, focused his research on fundamental questions concerning the nature of information, perception, and cognition. Dretske developed a detailed theory linking information and knowledge, known as the "semantic conception of information." In terms of the properties of information, Dretske primarily highlighted quantity and semantic aspects. Dretske proposed that the amount of information in a message can be measured depending on the extent to which it reduces uncertainty in decision-making by the information consumer, which in turn is identical to the main postulates of the Shannonian conception of information.

However, what was unique about Dretske's research was his postulation of the significance of truth and essentiality in information. For a data construct to be recognized as information in Dretske's understanding, it must not only be semantically significant but also true. False or misleading messages cannot be considered information in the full sense of the word, as they do not add to existing knowledge.

Furthermore, one of Dretske's key ideas was that a message or signal must embody a "claim to mattering" in the context of its interpretation. This means that information must be relevant and appropriate in a given context for it to be meaningful to an individual (Dretske, 1981).

The work of Karl Deutsch and Fred Dretske is the basis for this study. Deutsch emphasized the systemic impact of information on societal and political dynamics, focusing

on its ability to mobilize or suppress activity and redistribute power. His identification of properties such as the "tone" or "temperature" of information as measures of its emotional impact aligns with the study's focus on the qualitative dimensions of substantive-content messages. Similarly, Dretske's semantic theory of information, which links its value to the reduction of uncertainty and its truthfulness, underpins the quantitative aspects of this research. Dretske's emphasis on the relevance and contextual appropriateness of information informs the methodological design, particularly in modeling how intensity and frequency shape the substantive-content message's presence. By integrating these perspectives, this study advances a probabilistic framework that not only captures the complexity of information dynamics but also bridges the qualitative and quantitative dimensions identified in Deutsch and Dretske's seminal works.

A significant development of this idea can be seen in the research of Karl Pribram. Karl Pribram was a renowned neuropsychologist and a pioneer in the field of neuroscience, particularly in the study of memory, perception, and other cognitive processes. His contribution to the understanding of information can be viewed through the prism of his "holographic theory of memory", according to which information is encoded in the individual's brain in a distributed manner, similar to the nature of holography. In the context of his research, one can highlight such properties of information as its distributed nature, fractal complexity, interaction with other information, and its connection to context.

According to Pribram's holographic theory, information is not localized in a specific area of the brain but rather distributed throughout the entire volume of the brain, while also having a fractal nature, manifested in the fact that patterns of repetition and nesting are found at different levels of thought processing - from neural networks to higher cognitive functions. At the same time, Pribram postulated that information does not exist in a vacuum and it always interacts with other information. This interaction changes and redefines the significance of information depending on the context - similar to how an image in a hologram can change depending on the perspective from which it is illuminated. Information in Pribram's research is invariably connected to context. This property emphasizes that the brain's information processes do not simply process data but also create a substantive-content message based on experience and expectations (Pribram, 1981).

Therefore, it can be said that the nature of the human brain determines the substantive-content message of the data construct based on the information basis formed in the consumer's consciousness and in combination with the continuous context. Consequently, the very process of perception and the previously formed basis of this perception transforms the substantive-content message embedded by the source of information.

The scientific novelty of this study lies in the development of a probability distribution function for the presence of substantive-content message properties in the information background. This function integrates the intensity parameter, which reflects the depth and saturation of the substantive-content message, and the internal covariance parameter, which characterizes the relationships between the manifestations of these properties. Together, these parameters enable a detailed analytical analysis of the information background, capturing both the degree of importance of the topics and their frequency of occurrence. Current research on modeling information flows primarily focuses on either intensity or frequency, often treating these parameters separately and overlooking their interdependence. Moreover, existing methodologies tend to emphasize quantitative aspects while neglecting qualitative dimensions, such as thematic richness and the interrelations among components. These limitations reduce the ability of current models to explain the complex dependencies and dynamics inherent in multidimensional information

environments(Clark, D., Hunt, S., Malacaria, P. , 2007; Bruce, N. I. et al, 2017; Bashir, H. et al, 2022; Cappella, J. N., Li, Y., 2023; Lim, S., Schmälzle, R., 2023).

This study addresses these gaps by introducing an integrative approach that combines intensity and frequency parameters using gamma and multivariate normal distributions. This approach allows for a more comprehensive analysis, accounting for the depth, richness, and interrelationships of messages, thus providing a nuanced and accurate depiction of the information background. Additionally, the use of genetic algorithms to optimize distribution parameters enhances the model's flexibility and adaptability, opening new opportunities for the analysis and forecasting of information dynamics.

2. Methodology

The information background, presented as a collection of substantive-content messages of data constructs, can be represented in many forms at the physical level. The fundamental types in this case are symbolic and natural information. Natural information, as a category, unites data formed by natural phenomena and directly perceived by the individual's sensory organs (Goldberg, 2022; Nadkarni et al., 2011). This type of information includes both visual and auditory images, as well as complex structures that appear in natural languages formed during cultural and social development.

One of the key researchers in the analytical specificity of natural information is Leonard Euler, who proposed using graph models to describe systems presented in natural form (Stapleton et al., 2010). Also, a significant contribution to the study of the analytical specificity of natural information was made by Gregor Mendel in the framework of heredity research. Mendel's research results in the field of genetics allowed for a description of the mechanisms of understanding how information is encoded and transmitted at the biological level (Gliboff, 1999).

Symbolic information, in turn, is characterized by the use of signs and symbols of an artificial nature (created by individuals to represent and convey knowledge, ideas, and concepts). Language, mathematical notation, and computer codes are examples of symbolic systems.

A key specificity of symbolic information is the a priori need for interpretation, according to which each symbol or sequence of symbols is assigned a specific meaning in accordance with the regulations established within the cultural or disciplinary context. Symbolic systems are characterized by a high degree of abstraction and formalization, which allows for the use of logical and algorithmic tools for their analytical formalization (Onykiy et al., 2020; Beth, 2012).

The most significant research contribution in the field of symbolic information analysis was also made by Claude Shannon. The concept of entropy, formulated by Shannon, became a key tool for evaluating the amount of symbolic information and its transmission in digital form (Shannon, 1948). It is also necessary to recall the work of Alan Turing, primarily in the context of abstract computing devices capable of modeling algorithmic processes. Conditional "Turing Machines" most clearly demonstrate the concept of symbolic information, as they represent algorithms and data in a symbolic form (Turing, 1939).

Within the described methodology, the primary focus is on the universality of encoding the input array for comparison purposes. This leads to the postulation of the thesis that the final form of data constructs containing a substantive-content message requires a transition to a comparable symbolic form. Natural textual form is recommended as this form.

Natural textual form of information represents one way of encoding the substantive-content message using natural language. This language develops organically within social groups and cultures and differs from formal and artificial languages in its polysemy, its non-reducibility to a strict formal structure, and its flexibility in expressing individual elements and contexts of human experience. Natural textual form of information encompasses a wide range of variations in manifestation, including literary works, scientific documents, news articles, social media correspondence, legal documents, and much more. Textual constructs created by an individual are potentially saturated with a variety of semantic nuances, metaphors, allusions, and other syntactic figures, which in turn shape unique characteristics. Natural language has a complex hierarchy, levels of abstraction, and interacts with numerous levels of human cognition and social patterns of interaction (Li, 2024; Boerman et al., 2021; Hsieh et al., 2011).

The following components of information can be identified:

• Phonetic and phonological level: defining the acoustic characteristics of language.

• Morphological level: related to the structure and formation of lexemes from morphemes.

• Syntactic level: including systems of regulations and structures that determine the target combinations of lexemes in textual constructs.

• Semantic level: describing the deep structure of the substantive-content properties of lexemes, combinations of lexemes, and texts as a whole.

• Pragmatic level: focused on the use of language in the context of communication and the use of text in the implementation of social and cultural functions.

Moreover, the natural textual form of information has contextual, sociocultural, and cognitive features that determine how the text is perceived and interpreted by the individual. Aspects such as intonation, emotional coloring, and stylistic design largely shape the subjective component of the substantive-content message of the data construct. Thus, it is precisely the natural textual form of information that allows for a balance between the universality of analysis and substantive-content richness (Wang et al., 2023; Amur et al., 2023; Dehaene et al., 2011).

Methodologically, the primary stage in modeling the properties of the substantive-content message encoded in the form of symbolic data constructs is the collection and systematization of an array, which in this case is presented in natural textual form. The collection process can be differentiated according to two fundamentally different approaches - automated and non-automated (Biggers et al., 2023).

An automated approach implies interaction exclusively within the digital environment, which in turn leads to a dualism of the final data array. On the one hand, the information aggregated within the approach is secondary, which in turn increases the complexity of specifying the target properties of the substantive-content message. However, the dissociation of the subject of generation from the subject of aggregation allows for the primacy of the substantive-content message in relation to the impact that inevitably occurs during communication between these subjects (Weinlich et al., 2022).

Technologies for automated information gathering in the digital environment are described in the direction of parsing. Parsing is the process of automated extraction of information aggregated by web resources. The concept includes numerous operations, such as requesting web resource content and extracting and structuring data, which allows

transforming information presented in various forms, particularly in natural textual form (Lytras et al., 2020).

The following methodological stages of parsing can be identified:

• Data collection: In the first stage, a browser or other software tools for automated query sending are used to extract the contents of the web resource in the form of HTML, JavaScript, and other web technologies.

• Scraping: This stage involves the analysis of the extracted HTML code of the web resource, which includes parsing the structure of the DOM (Document Object Model) of the resource. The task of scraping is to determine the data structure and extract the necessary elements (for example, text, links, images, etc.).

• Data extraction: The next step involves selecting specific data from HTML elements based on their classes, identifiers, or other attributes, which can be done using XPath, CSS selectors, or other mechanisms to designate paths to data in the DOM structure.

• Data processing: Based on the results of extraction, data can be cleaned, normalized, and transformed. This process includes removing redundant or undesirable symbols, converting formats (for example, dates and times), breaking down text blocks into more detailed elements, or organizing data into tabular formats.

The algorithm described above is quite generalized and is significantly differentiated depending on the specifics of the web resource. Conceptually, three key variations of the parsing approach can be identified: API parsing, web parsing, and simulation parsing. Web parsing is methodologically fully described by the provided algorithm. API parsing, on the other hand, is significantly less labor-intensive, as web resources equipped with APIs are initially created for parsing tasks, essentially enabling the complete exclusion of data extraction and processing stages. Simulation parsing, in turn, is significantly more labor-intensive to use, which in turn is due to the multidimensionality, interactivity, and differentiation of the web resource, requiring the simulation of individual actions to interact with the web resource.

A non-automated approach is significantly more instrumentally differentiated and involves data collection methods such as observation, surveys, questionnaires, focus groups, and much more. The result of implementing the collection stage is a structured array of symbolic data constructs.

The data includes natural language elements like punctuation, grammar, and simple words. This specificity defines the need for rectification of aggregated symbolic data constructs. A highly saturated and universal algorithm for rectifying natural textual information contains the following sequential steps:

• Tokenization: This is the process of dividing text into elementary units called tokens. Tokens can be words, phrases, symbols, or other elements into which text can be divided for subsequent analysis. It is the tokens that act as descriptive units of the substantive-content message of the data construct.

• Register universalization: This is the process of converting all alphabetic characters in the text to a single register (upper or lower). This step is necessary within the framework of rectifying natural textual information, as the register is a technical property of the token.

• Morphological tagging: Also known as part-of-speech tagging or POS-tagging, this is the process of assigning parts of speech, such as nouns, verbs, adjectives, and other

grammatical categories, to each token in the array. This procedure is necessary for subsequent filtering and processing of tokens.

• Primary filtering: Excluding tokens that do not meet the target properties, particularly those corresponding to parts of speech. The substantive specificity of filtering is largely determined by the research objectives.

• Lemmatization: This is a linguistic process of reducing word forms to their base form, known as the lemma. The lemma constitutes the main, dictionary, or original spelling of a word, which serves as its canonical form. It is at the stage of lemmatization that tokens are maximally universalized from a technical standpoint, which in turn allows for achieving maximum accuracy in identifying the substantive-content message.

• Secondary filtering: This involves removing individual tokens or categories of tokens from the array. This is a variable step and assumes the preliminary formation of an array of conditionally undesirable tokens, which can primarily include a priori thematic and low-content tokens.

As a result of this stage, a structured array of rectified symbolic data constructs is formed. For the purposes of subsequent modeling, data constructs presented as arrays of tokens need to be represented numerically, for which vector structures are most suitable. Vectorization provides a transformation of textual information into numerical vectors, suitable for analytical work and the execution of machine learning algorithms.

Within the framework of vectorization, a vector space is constructed where each document, sentence, or phrase is represented as a vector of numerical features. This operation allows the transformation of textual data into a structured form, enabling the identification of patterns and dependencies that are not obvious when analyzing the original text. The choice of vectorization method depends on the specific goals and tasks of data analysis. Key vectorization methods include One-Hot Encoding, Bag of Words, and Term Frequency-Inverse Document Frequency (TF-IDF).

One-Hot Encoding in the context of text vectorization is a process where categorical data of words is transformed into vectors of a fixed length, where each unique element of the text corresponds to a single active unit in the vector, and the rest of the positions are equivalent to zeros. As a result of this encoding, each data construct in the form of text is represented by a sparse vector where the tokens of the basic dictionary are equal to 1, and the vector dimension corresponds to the size of the dictionary. One-Hot Encoding has the advantage of simplicity of implementation and explicit indication of the presence of a token in the text, however, it has a number of disadvantages: vectors are formed extremely sparsely, which leads to a lack of consideration for the semantic proximity between tokens and thus reduces the informativeness of the future model. Therefore, the method is most effective in tasks where deep semantic analysis is not required and the dictionary size remains relatively small.

The Bag of Words model, in turn, is also a methodological approach in the field of natural language processing, the main assumption of which is the representation of text as a multiset of its constituent tokens without taking into account grammatical structure and syntax, as well as the order of tokens in the document. At the same time, this model can be presented as a way to transform textual data into numerical vectors based on the frequency of occurrence of a token. Accordingly, this model represents each document from the corpus as a vector in N-dimensional space, where N is the number of unique tokens in the entire corpus of texts. Implementation of this model leads to the creation of a term-

document matrix, where rows correspond to individual documents of the corpus, and columns correspond to unique tokens of the dictionary - thus, each element of this matrix indicates how many times a token appears in a specific document. The key disadvantages of this method include, in particular, the sparsity of the resulting vectors with a significant dictionary and the loss of information about the order of tokens, which also leads to the inability to account for contextual and syntactic relationships in the model. The method also does not take into account the significance of words - all tokens are considered equally significant, although some of them contain more substantive-content message.

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure for assessing the importance of a token within the context of a document that is part of a corpus of texts. This method is based on two main metrics: term frequency (TF) and inverse document frequency (IDF). Text vectorization using TF-IDF allows shifting the focus away from the usual frequency of occurrence of tokens in a document to those that are more unique and potentially more saturated with substantive-content message.

The application of TF-IDF in the context of text vectorization allows obtaining a vector representation that demonstrates tokens that are significant precisely within the framework of a particular document within the context of the corpus, thereby reducing the significance of those tokens that frequently appear in other documents of the corpus. However, one of the disadvantages of this method is that TF-IDF can also reduce the significance of high-frequency, but at the same time, content-rich tokens if they appear in a large number of documents in the corpus. Also, this method does not solve the problem of contextual significance and does not take into account the order of tokens. Based on the described specifics, TF-IDF is most appropriate to use in those automatic text processing systems where it is necessary to identify and filter key terms and deep understanding of the substantive-content load of the text is not required.

Word Embedding is a vectorization method that represents a dense and continuous vector space where each token in the corpus of texts is mapped as a point in space. These vectors are formed in such a way that the contextual semantics of tokens are taken into account, and tokens with similar substantive-content messages are located closer to each other in multidimensional space. The advantages of Word Embedding lie in its ability to reflect semantic relationships and the diversity of words depending on context, as well as in a significant reduction in dimensionality compared to traditional approaches, such as one-hot encoding. However, a significant disadvantage is that the quality of embeddings largely depends on the volume and quality of the training sample, as well as being relatively labor-intensive to interpret high-dimensional vectors.

The result of implementing one of the described methods is the formation of a set of vectors describing the lexical content of symbolic data constructs, which in turn acts as a quantitative array for training a model to evaluate the relative presence of substantive-content message properties.

Within the stage of constructing this model, it is necessary to consider two fundamental scenarios: whether the obtained array is labeled in an endogenous context or not, and therefore, whether it is necessary to use supervised or unsupervised learning methods. In supervised learning tasks, data arrays include both input and output variables. Input variables represent a feature matrix where each vector describes one example from the dataset, in particular, one document in the corpus of texts. Each example is described by a set of characteristics, which can be obtained through various text vectorization methods. In turn, output variables contain class labels or values that are the target of model predictions.

These two arrays are used synergistically to train the model so that it can identify dependencies between input and output data.

Machine learning algorithms in the context of topic modeling based on vectorized texts can be divided into several main categories: linear, logical, ensemble, and deep learning methods, each of which has specific features of application in topic modeling tasks.

• Linear methods in topic modeling are based on the assumption of linear separability in the data. In the context of vectorized texts, they strive to form a separation boundary that maximally distinguishes texts of different topics. However, actual data often contains not only linear dependencies, especially in the case of complex semantic distribution.

• Logical methods, in particular decision trees, implement hierarchical separation of the feature space, striving for maximum purity of nodes relative to the thematic affiliation of documents. These methods are effectively interpretable, but prone to overfitting and unstable to changes in the data.

• Ensemble methods, such as Random Forest or gradient boosting, combine multiple simpler models (e.g., decision trees) for the purpose of improving the predictive power and stability of the result of application. In topic modeling, they can capture complex structures in the data but are difficult to interpret.

• Deep learning methods include neural network architectures capable of detecting complex non-linear relationships in data. Applicable to topic modeling based on vectorized texts, significant architectures include convolutional neural networks (CNNs) for analyzing local word relationships and recurrent neural networks (RNNs), as well as their variants, in particular, LSTM and GRU, which function effectively when processing sequential data and can capture the contextual dependence of tokens in text. Deep learning is effective for identifying deep semantic content, however, it requires large amounts of data for quality training and significant computational resources, and model architectures are exceptionally non-interpretable.

In unsupervised learning tasks, the data array consists only of input variables, as the goal of this type of learning is to study the structures and patterns in the data without any predefined labels. The input also uses a feature matrix created based on text vectorization, but in the absence of an endogenous variable, training takes place based on the statistical structure of the data.

Unsupervised learning algorithms, such as clustering or dimensionality reduction, explore the input matrix to identify internal connections and patterns, such as latent factors, characteristic data groups, or principal components. To solve these problems within the framework of topic modeling, clustering methods are often used, such as LDA, K-Means, Hierarchical Clustering, DBSCAN, Spectral Clustering, and Mean-Shift Clustering.

• Latent Dirichlet Allocation (LDA) is a generative statistical model that allows data sets to express explanations over observed documents. In LDA, each document is represented as a random mixture of latent substantive-content messages, and each cluster is represented as a hierarchical mixture of tokens.

• K-Means is an algorithm that partitions n observations into k clusters, minimizing intra-cluster variations and maximizing inter-cluster differences. Although K-Means can be effective in processing large arrays, it assumes that clusters are convergent and have the same volume and density, which can be a limitation when topic modeling.

• Hierarchical clustering creates a cluster tree, which can be useful for detecting nested cluster structure, but may be inapplicable to large arrays due to high computational costs.

• DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that identifies areas of high density separated by areas of low density. DBSCAN effectively works with clusters of arbitrary shapes and can identify outliers in the data.

• Spectral clustering uses the eigenvalues of the data dimension to project them into a smaller space before performing clustering in the new space. This method is effective for complex structures, but computational complexity and sensitivity to scaling can also be obstacles when working with large arrays.

• Mean-Shift Clustering is based on finding dense areas of data. The model does not require pre-determining the number of clusters, but it can be ineffective in cases of high data dimensionality.

Each of the methods listed has its own advantages and disadvantages in the context of topic modeling. Choosing the right method depends on the size and structure of the array, the desired resolution of the thematic structure, as well as on computational resources and the required level of interpretability of the results. LDA is the most versatile and is defined by deep topic modeling, while other methods may be preferable in the framework of fast pre-processing tasks.

Therefore, the main difference between arrays in supervised and unsupervised learning tasks lies in the presence or absence of an output array with labels, which is necessary for mapping input data to a specific desired outcome in controlled tasks and absent in uncontrolled tasks, where only relationships between input data are analyzed.

As a result of implementing this stage, a model is formed to evaluate the relative presence of substantive-content message properties. In the final stage of the developed methodology, an analysis of the qualitative properties of the obtained model is performed. Quality metrics differ significantly depending on the type of model - classification, regression, or clustering. The main quality metrics of classification models include the Confusion Matrix, Accuracy, Precision, Recall, and ROC curve.

• Confusion Matrix is a table that compares the actual class labels to the labels predicted by the classification model. The matrix consists of four parts: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) classifications. In the context of text classification, this matrix allows us to assess how often the model correctly or incorrectly classifies documents of each class.

• Accuracy is the most intuitive metric and is calculated as the ratio of the sum of true positives and true negatives to the total number of elements in the sample. However, if the classes are unbalanced, high accuracy may not reflect the true quality of the classification.

• Precision also allows assessing the accuracy of model predictions, namely the proportion of true positives among all those considered by the model as positive. In the context of text analysis, this metric allows us to determine what proportion of documents labeled by the model as belonging to a certain category actually belong to that category.

• Recall reflects the model's ability to identify all relevant cases in the data, that is, the proportion of true positives relative to all elements of that class. For text classification, this metric indicates what proportion of documents of a particular topic the model was able to correctly identify.

• Receiver Operating Characteristic (ROC) curve and the corresponding AUC (Area Under the ROC Curve) metric are used to measure the quality of binary classifiers. The ROC curve is plotted with True Positive Rate (TPR, synonymous with Recall) against False Positive Rate (FPR). AUC represents the area under the ROC curve and can serve as an aggregated measure of model performance across all classification thresholds. In the context of text classification, the ROC curve allows us to assess the quality of text separation into two categories regardless of the limit threshold.

The quality metrics of regression models provide a means to assess the level of continuous prediction accuracy relative to actual values. The coefficient of determination (R^2) measures the proportion of variation in the dependent variable that is explained by the independent variables in the regression model. In the context of text data analysis, R^2 reflects how well the vectorized features of texts predict the target numerical attribute.

The Mean Absolute Error (MAE) is calculated as the average of the absolute differences between predicted and actual values. MAE provides insight into the average magnitude of the model's errors without considering their direction. In the context of text regression, MAE can show how far predictions, such as sentiment scores, deviate on average from the actual ratings.

The Mean Absolute Percentage Error (MAPE) incorporates prediction errors as percentages of actual values, which allows for interpreting the size of errors in the scale of the target variable. However, MAPE should not be used if there are values close to zero in the data, as this can cause uncertainty due to division by small numbers.

The Mean Squared Error (MSE) measures the average squared difference between predicted and actual values. This metric is more sensitive to outliers. In text regression tasks, MSE can be useful for identifying cases where the model produces several extremely significant deviations, even if most predictions are accurate.

When interpreting results, low values of MAE, MAPE, and MSE indicate higher quality predictions from the model, while a high coefficient of determination demonstrates that the model effectively explains the variability in the data. In the context of regression models for text data, it is important to consider how dispersed the prediction errors are. If the model systematically underestimates or overestimates certain observations, the metrics may become skewed (Weaver, 2017).

In the context of thematic clustering of texts, where it is important to obtain groups of documents with a homogeneous theme, quality metrics such as Silhouette Score, Calinski-Harabasz Index, and WCSS (Within-Cluster Sum of Squares) are used.

• Silhouette Score measures the quality of clustering based on how close each object is to objects of its cluster compared to objects of other clusters. The silhouette score ranges from -1 to +1, where a high positive score indicates that the object is effectively related to its own cluster and ineffectively related to neighboring clusters. If the silhouette score is sufficiently high in the context of text clustering, then it can be argued that the clusters are effectively separated thematically and concentrated.

• Calinski-Harabasz Index is another metric for measuring clustering quality, based on the ratio of the sum of distances between clusters to the sum of intra-cluster distances. High values of this index correspond to denser and more separate clusters, which is desirable when clustering. Applied to texts, a model with a high Calinski-Harabasz index value typically forms tightly clustered clusters with clear thematic boundaries.

• WCSS (Within-Cluster Sum of Squares), or the sum of squared intra-cluster distances, reflects the sum of squared distances from each object to the center of its cluster. This metric is used to determine the optimal number of clusters, when a value of k is sought where increasing the number of clusters does not lead to a significant reduction in WCSS, which can be seen on the corresponding elbow plot. In the context of texts, a low WCSS value indicates that documents within clusters are thematically similar to each other.

Each of the metrics presented is not universal. The silhouette is most effective for assessing the direct quality of clustering, it gives an understanding of how isolated the clusters are. The Calinski-Harabasz index is effective for comparing the quality of different cluster models, especially when it is necessary to assess the influence of the number of clusters on the overall structure. WCSS is useful in determining the number of clusters, but it can only be applied if inter-cluster distances are comparable.

As a result of this stage, a set of quality metrics for the model to assess the relative presence of substantive-content message properties is formed. Appendix shows the described methodology for modeling the properties of the substantive-content message encoded in the form of symbolic data constructs.

The diagram at Appendix presents a structured, step-by-step methodology for analyzing the relative presence of properties of the essential-contentual message in a data environment, divided into five stages. The process begins with data collection, employing either automated approaches (e.g., API parsing, web scraping, or simulated parsing) or manual techniques (e.g., NLP and string processing), resulting in a structured array of character data constructs. In the data preprocessing stage, the raw data is refined through steps such as tokenization, case normalization, morphological tagging, filtering, and lemmatization, producing a rectified data array ready for vectorization. The vectorization stage transforms the processed data into numerical representations using methods like one-hot encoding, bag of words, TF-IDF, or word embeddings, generating a complex of vectors that encapsulate the lexical content. These vectors are then analyzed in the modeling stage, using supervised learning tasks (e.g., linear models, logistic regression, ensemble models, and deep learning) or unsupervised clustering algorithms (e.g., LDA, K-Means, DBSCAN) to evaluate the presence of essential-contentual message properties, producing a model for analysis. Finally, in the quality analysis stage, the models are evaluated using metrics like accuracy, precision, recall, R^2, MAE, and clustering scores (e.g., Silhouette Score, WCSS), ensuring reliability and effectiveness. This methodology provides a comprehensive framework for processing, modeling, and assessing data, culminating in actionable insights into the distribution and properties of the essential-contentual message.

The developed algorithm for multimodal analysis integrates intensity modeling (using the Gamma distribution) and frequency modeling (using the multivariate normal distribution) to capture the depth and recurrence of substantive-content message properties. These two components are combined into a unified probabilistic framework, optimized via a genetic algorithm to identify the most accurate parameters. The resulting optimized model provides a foundation for generating analytical insights and practical applications, enabling the exploration of complex thematic relationships within the information background. The visual representation of the algorithm is provided in Figure 1, offering a clear, structured view of the methodology's operational flow and its integration of key mathematical and analytical components.

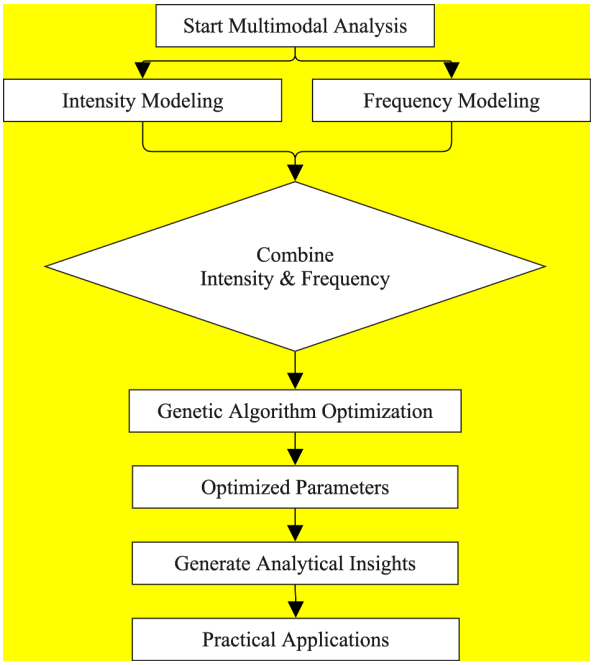

Figure 1. Multimodal Probabilistic Analysis Algorithm for Substantive-Content Message Properties

3. Results and Discussion

The distribution of the presence of substantive-content message properties in the information background is multimodal, and the specifics of the correlation of peaks are the most content-significant from the point of view of analysis. To describe this distribution, it is proposed to represent it as a product of two components: intensity and frequency of presence of substantive-content message properties in the information background.

Intensity reflects how strong or detailed an information signal is. In this case, we are not talking about the mere presence of a particular topic or narrative in the information background, but about the measure of its influence, significance, or emotional saturation. Intensity can manifest through the detail of discussion, the degree of emotional involvement of the audience, or the depth of analytical coverage of the topic.

The frequency of presence of substantive-content message properties, on the other hand, refers to the number of times that specific information or a topic manifests in the information background in a limited time frame. It measures how often the entity manifests in messages, publications, or discussions, regardless of the depth or context of its presence. In terms of probability and statistics, frequency can be associated with a series of events in the Poisson process, where the interest lies in the appearance of specific topics in the information background.

The key difference between intensity and frequency lies in quality versus quantity: intensity focuses on the quality of the presence of information (how deeply or richly the topic is presented), while frequency measures the number of times a topic appears, regardless of context or depth of manifestation.

For the purposes of approximating the described properties, it is proposed to use separate types of distributions. To approximate the intensity of presence of substantive-

content message properties in the information background, it is proposed to use the Gamma distribution.

The Gamma distribution, denoted as G (k, θ), where k > 0 is the shape parameter and θ > 0 is the scale parameter, is a continuous distribution defined on the positive semi-axis. The probability density of the Gamma distribution is given by the function (see equation 1):

$$f(x; k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}, \; x > 0 \qquad (1)$$

where G(k) is the Gamma function, defined as (see equation 2):

$$G(k) = \int_0^{\infty} x^{(k-1)} e^{(-x)dx} \qquad (2)$$

In the context of analyzing the information background, the Gamma distribution stands out for its adaptability and ability to model diverse phenomena associated with the accumulation of events, making it particularly relevant for the quantitative study of the intensity of the presence of the substantive-content message. When analyzing the information background, the Gamma distribution can be used as a mathematical model to estimate the parameters of the distribution of events over time. In this context, k is interpreted as the expected number of "events" (in this case, the appearance of specific themes or ideas), and θ is a measure of time or space during or within which these events can occur. This in turn forms the basis for probabilistic modeling and statistical analysis of the intensity of presence of thematic narratives in the information background, allowing us to assess both the probability of a specific intensity of discussion and calculate the expected indicators of the appearance of informational entities.

To approximate the frequency of the presence of substantive-content message properties in the information background, it is proposed to use the multivariate (or multidimensional) normal distribution.

The multivariate (or multidimensional) normal distribution is a generalization of the univariate normal distribution for a vector of n variables. The multivariate normal distribution is defined by a vector of means (μ) and a covariance matrix Σ, where each element of the vector (μ) represents the mean of the corresponding variable, and each element of Σ represents the covariance between the variables.

The probability density for the multivariate normal distribution is given by the expression (see equation 3):

$$f(x; \mu, \Sigma) = 1 / \left((2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}} \right) \exp\left(-\frac{1}{2(x-\mu)^T \Sigma^{-1} (x-\mu)} \right) \qquad (3)$$

where **x** — is the vector of variables, n — is the dimension of the variable space, |Σ| — is the determinant of the covariance matrix, and Σ-1 is the inverse covariance matrix.

The multivariate normal distribution is effective for modeling and analyzing the frequency of the presence of the substantive-content message in the information background as it enables the capture and modeling of correlations between various entities in the information space. This allows for analyzing how the appearance of one topic in the information background is related to the appearance of other topics. Also, due to the covariance matrix, the multivariate distribution can adapt to different types of relationships

between variables, including uncorrelated, positively and negatively correlated data, allowing for an accurate description of complex relationships between the frequencies of appearance of various topics. In the modern world, information often has a multidimensional nature (e.g., text data, metadata, context-dependent properties). The multivariate normal distribution can effectively model and analyze this multidimensionality, providing valuable insights into the structure of information interactions. Therefore, using the multivariate normal distribution to describe the frequency of the presence of the substantive-content message in the information background allows for a deeper understanding and quantification of the relationships and structure of information flows. At the same time, the covariance matrix can be represented by a uniform parameter, which meaningfully reflects the probability of dependence of nearby values in the array on each other. Thus, the distribution of the presence of substantive-content message properties in the information background can be described by the following parameters:

k – The shape of the intensity of the presence of substantive-content message properties in the information background. In the context of the Gamma distribution, the shape parameter allows characterizing the following specifics of the analyzed properties:

• Level of concentration of the distribution. Low values of the shape parameter indicate a more concentrated distribution, while higher values are characteristic of more distributed data. This specificity may indicate the significance of the diversity of the intensity of the presence of the analyzed substantive-content message property in the information background.

• Level of homogeneity of the distribution. When analyzing the information background, the shape parameter allows for the formulation of comparative conclusions about how diverse or homogeneous the events related to the appearance of the analyzed substantive-content message properties are. A higher level of this parameter indicates greater diversity, where events are more evenly distributed over time or information space.

• Level of variability of events. The variability of the data relative to the average value increases with an increase in the shape parameter, which in turn indicates a more predictive presence of the substantive-content message properties with higher values of k.

The Gamma distribution was selected to model the intensity of substantive-content message properties because of its capacity to represent positively skewed phenomena, which are characteristic of the depth and saturation of information signals. Its shape parameter (k) captures the concentration and variability of intensity, while the scale parameter (θ) accounts for the range of distribution, making the model adaptable to various contexts. Meanwhile, the multivariate normal distribution was chosen for modeling the frequency of these properties due to its ability to describe correlations and co-occurrences among multiple variables. By using a mean vector (μ) and covariance matrix (Σ), it provides a detailed depiction of thematic relationships and multidimensional data structures. Together, these distributions complement each other, allowing for a comprehensive representation of both the quality (intensity) and quantity (frequency) aspects of the information background. This integration forms a robust probabilistic framework capable of accurately modeling multimodal data, enabling advanced analysis and predictive insights.

General specificity of the information background slice or the analyzed information flow. The probabilistic model, including the shape parameter, can reflect the specificity of the information flow, in particular, how often and with what intensity certain properties

appear. In particular, in the media space, characterized by intense thematic renewal, the shape parameter will allow us to draw conclusions about the level of relevance or saturation of the information background.
θ – The scale of the intensity of the presence of substantive-content message properties in the information background. In the context of the Gamma distribution, the scale parameter allows us to characterize the following specifics of the analyzed properties:

• Intensity of targeted information flows. A high value of θ indicates a wider spread of data, which may indicate a high intensity of targeted information flows with a more diverse frequency of occurrence. This fact indicates that the substantive-content message may appear frequently in some periods and significantly less frequently in others, providing a non-uniformly distributed pattern.

• Range of appearance. The parameter θ also allows us to assess the variability of the appearance of informational entities, i.e., how often and in what volumes the information message is found in the information background.
μ - The vector of mean values, reflecting the central tendency of the presence of substantive-content message properties in the information background. In an analytical sense, the parameter μ indicates the dominant position of substantive-content message properties within the considered multidimensional space. Interpretation of the μ value allows identifying the general orientation of the information field and assessing around which axes of essential or content parameters the data is grouped, thereby reflecting the average statistical characteristic of the prevailing trend in the extensive information flow. Thus, the value of μ in this analytical context is an indicator of the midpoint of the attributes of the substantive-content message, from which one can identify both general trends in the information space and transitions in the dominance of certain properties.

Σ – The covariance matrix, reflecting the degree of interrelationship between substantive-content message properties in the information background. The elements of this matrix indicate the covariance between each pair of variables included in the analysis. In the study of the distribution of the presence of substantive-content message properties in the information background, this parameter indicates the structure and dynamics of the interaction of these properties. When analyzing the information background based on the multivariate normal distribution, Σ allows us to assess how synchronously the discussed topics, emotional coloring, and other substantive-content aspects change.

Each of the identified parameters has significant analytical specificity. However, parameters k and μ meaningfully reflect unified characteristics of the presence of substantive-content message properties in the information background, which allows us to equate them to each other. Parameter θ, in turn, is less analytically significant, as the comparison of distributions in this case requires scaling, which in turn implies normalization. Therefore, parameter θ can be equated to 1.

The covariance matrix Σ, on the other hand, has a more specific nature. Since the distribution of each substantive-content message property is considered separately from the others, interrelationships can be considered solely within the framework of the elements of a single array. Thus, the covariance matrix can be replaced by a homogeneous scalar-augmented matrix, all elements on the main diagonal of which are equal to 1, and all off-diagonal elements are equal to each other and not equal to the diagonal elements. Off-diagonal elements, in turn, reflect the level of dependence of the manifestation of properties of adjacent elements of the array. As it increases, the probability of manifestation of the property in elements adjacent to the dominant one increases. This parameter can be

denoted as the coefficient of internal covariance, and the matrix formed based on it will take the form (see equation 4):

$$\Sigma_{ij} = \begin{cases} 1 \text{ if } i = j \\ Corr_i \text{ if } i \neq j \end{cases} \quad (4)$$

Thus, the distribution of the presence of substantive-content message properties in the information background can be described by two conceptual parameters:

• k: The intensity parameter, reflecting the depth and saturation of the substantive-content message.

• Corr: The internal covariance parameter, characterizing the relationship between the manifestations of the properties of the substantive-content message.

To identify these parameters, it is necessary to approximate the distribution of the data array describing a particular property of the substantive-content message. However, the simultaneous use of continuous and discrete distributions does not allow using traditional methods such as the method of moments or the method of maximum likelihood. For the purposes of iterative data parameter selection, genetic algorithms can be used, as:

Genetic algorithms are useful for exploring many parameter combinations in complex models.

• Genetic algorithms are able to identify the global optimum of a function, even if the optimization landscape contains many local minima and maxima.

• Thanks to their adaptability and flexibility in tuning, genetic algorithms can be effectively adapted to the task of finding the best parameters for functions describing the distribution of data with complex interdependencies between variables.

• Unlike traditional optimization methods, which require formal definition of gradients or other function derivatives, genetic algorithms can optimize parameters without the need to define the exact form of the distribution.

These properties indicate the potential effectiveness of using genetic algorithms in the context of the task at hand.

In the first stage of the algorithm, an initial population is formed. The initial population in this case is represented by arrays of values of the internal covariance coefficient of the manifestation of substantive-content message properties in the information background (Corr) and values of the intensity coefficient of the presence of substantive-content message properties in the information background (k) within specified ranges. As ranges of coefficients, values from 0 to 1 (exclusive) can be specified. The size of the initial population is also variable and can be denoted as P.S. (see equations 5,6).

$$Corr = \{Corr_i | Corr_i \text{ simU}(0.01,\ 0.99),\ i = 1,\ 2,\ \ldots,\ P.S.\} \quad (5)$$

$$k = \{k_i | k_i \text{ simU}(0.01,\ 0.99),\ i = 1,\ 2,\ \ldots,\ P.S.\} \quad (6)$$

Where:

$Corr_i$ – The i-th variation of the internal covariance coefficient of the manifestation of substantive-content message properties in the information background.

k_i – The i-th variation of the intensity coefficient of the presence of substantive-content message properties in the information background.

P. S. – The size of the initial population.

Corr – The generated array of values of the internal covariance coefficient of the manifestation of substantive-content message properties in the information background.

k – The generated array of values of the intensity coefficient of the presence of substantive-content message properties in the information background.

Based on the pairs of generated values, a simulation model is constructed, differentiated into multivariate normal and gamma distributions. Let us first consider the modeling of the multivariate normal distribution. The first stage of this process involves generating a sample from the multidimensional normal distribution (see equation 7):

$$Y = [y_1, \ \ldots, \ y_{NSim}], \ where \ y_i \sim \frac{1}{\left((2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}\right)\exp\left(-\frac{1}{2(x-k)^T\Sigma^{-1}(x-k)}\right)}, \ \Sigma_{ij} = \begin{cases} 1 \text{ if } i = j \\ Corr_i \text{ if } i \neq j \end{cases} \quad (7)$$

Where:

Y – A matrix of random vectors of size NSim × n.

y_i – The i-th row vector in the matrix Y, representing a single simulation from the distribution.

NSim – The number of simulations.

Σ_{ij} – A homogeneous scalar-augmented matrix, all elements on the main diagonal of which are equal to 1, and all off-diagonal elements are equal to $Corr_i$.

Next, the quantile of the standard normal distribution is determined, and each element of the sample is compared to the quantile, and the result is converted to 0 or 1 (see equation 8):

$$Q = \left[\left(y_i > F^{-1}(1 - k_i)\right)\right] \quad (8)$$

Where F^{-1} denotes the inverse distribution function (quantile function) of the standard normal distribution, and the intensity coefficient of the presence of substantive-content message properties in the information background (k_i) indicates the specified probability level. Each element y_i of the matrix Y is compared to the quantile, and the result of the comparison is transformed: if y_i is greater than the value of the quantile, then 1 is placed in Q, and 0 otherwise. This transformation is applied element-wise for each value in Y.

Finally, the matrix of results is transposed and its type is converted (see equation 9):

$$Q' = \text{transpose}(Q) \quad (9)$$

Where Q′ is the transposed matrix of Q. The matrix formed as a result describes the frequency of manifestation of substantive-content message properties in the information background. To account for intensity, a corresponding simulation of the Gamma distribution is necessary. Generation of the matrix of random intensity values is determined by the equation 10.

$$M = [M_{ij}], \text{ where } M_{ij} \sim \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)}, \ \theta = 1 \quad (10)$$

Each element M_{ij} of the matrix M is a random number generated from a Gamma distribution with parameters k_i and θ = 1, thus forming a random matrix of dimension NSim × NSim. The resulting arrays are then multiplied and the results normalized (see equations 11,12):

$$S = \sum_{i=1}^{n} Q'_i \cdot M_i \quad (11)$$

$$Z = \left\{\frac{s_1 - q_S}{\sigma_S}, \frac{s_2 - q_S}{\sigma_S}, \ \ldots, \frac{s_N - q_S}{\sigma_S}\right\}, \text{ where } \ q_S = \frac{1}{n}\sum_{i=1}^{n} s_i, \ \sigma_S = \sqrt{\sum_{i=1}^{n}(s_i - q_S)^2} \quad (12)$$

Where:

Z – The resulting simulation array.

s_i – The i-th value of the degree of presence of substantive-content message properties in the information background.

q_S – The average value of the degree of presence of substantive-content message properties in the information background.

Each of the resulting arrays Z represents a variation of the distribution of the degree of presence of substantive-content message properties in the information background. The properties of the distribution of the generated arrays Z are correlated with the initial array R, describing the presence of substantive-content message properties in the information background of the enterprise. The comparison is made based on the Mean Absolute Error (MAE) between two rank vectors obtained from the distribution histograms. The number of bins in the histogram determines the level of abstraction of the form and is expressed by the discretization coefficient DC. Thus, the histogram for the data set Z with the number of bins DC, where each bin is denoted as B_p for k = 1, 2, ..., DC. The frequency of data entering each bin B_p, denoted as $C_Z(B_p)$, and defined as (see equation 13):

$$C_Z(B_p) = \sum_{i=1}^{n} 1(z_i \in B_p), \ C_R(B_p) = \sum_{i=1}^{n} 1(r_i \in B_p), \ p = 1, \ 2, \ \ldots, \ DC \quad (13)$$

Where $1(z_i \in B_p)$ — is the indicator function, which is equal to 1, if z_i is within the range of bin B_p, and 0 otherwise. After the frequencies for all bins $C_Z(B_p)$ are calculated, the ranks of each bin can be determined. Let C_Z represent the sequence of frequencies for the bins. First, a sequence is created that results from sorting C_Z in ascending order. Denote this action as $S(C_Z)$, where each element $C_Z(B_p)$ is assigned its index in the sorted sequence. Then, the rank of each bin $R_Z(B_p)$ is defined as the index of its frequency $C_Z(B_p)$ in $S(C_Z)$. Thus, the bin with the lowest frequency is assigned the lowest rank, and so on sequentially for all other frequencies as they increase (see equation 14).

$$R_Z(B_p) = \text{index}\left(C_Z(B_p) \text{ in } S(C_Z)\right), \ R_R(B_p) = \text{index}\left(C_R(B_p) \text{ in } S(C_R)\right) \quad (14)$$

This process of calculating the rank of each bin in the histogram allows for the formation of comparable sequences, based on which the MAE is then calculated (see equation 15).

$$MAE = \frac{1}{DC}\sum_{k=1}^{DC} |R_Z(B_p) - R_R(B_p)| \quad (15)$$

Thus:

$C_Z(B_p)$ – The frequency of data entering each bin B_p for the generative set Z.

$C_R(B_p)$ – The frequency of data entering each bin B_p for the actual set R.

$1(z_i \in B_p)$ – The indicator function, equal to 1 if z_i is within the range of bin B_p, and 0 otherwise.

B_p – A specific bin of the histogram.

DC – The discretization coefficient, determining the number of bins in the histogram.

$S(C_Z)$ – The sequence C_Z sorted in ascending order.

$S(C_R)$ – The sequence C_R sorted in ascending order.

$R_R(B_p)$ – The rank of each bin in the histogram of the actual set R.

$R_Z(B_p)$ – The rank of each bin in the histogram of the generative set Z.

MAE – The Mean Absolute Error between the two rank vectors obtained from the histograms of distributions Z and R.

The formed array of MAE values allows ranking the generated pairs of Corr and k by the level of approximation quality. The half of the most effective pairs form the first part of the new population. The second part of the new population is formulated based on the mutation of the first, implying a change in each of the values of the first part within a range of 20% with a probability of 50% (see equations 16).

$$F.C._i = \left\{(Corr_i,\ k_i)\middle|MAE_i < \left(\frac{1}{P.S.}\sum_{j=1}^{P.S.} MAE_j\right)\right\},\quad \forall i \in \{1,2,\dots,P.S.\}$$

$$(Corr_i', k_i') = \begin{cases} (Corr_i, k_i), & Prob = 0.5 \\ (Corr_i \cdot (1 \pm rand[0,0.2]),\ k_i \cdot (1 \pm rand[0,0.2])), & Prob = 0.5 \end{cases}$$

$$Corr = [Corr_i, \dots, Corr_m, Corr_i', \dots, Corr_m']$$

$$k = [k, \dots, k_m, k_i', \dots, k_m'] \qquad (16)$$

The updated population is redirected to the simulation modeling stage. This process is carried out for the required number of generations (K.T.), as a result of which the pair Corr and k with the lowest MAE value is identified (see equation 17):

$$(Corr^*, k^*) = \min_{(Corr_i,\ k_i)\in F.C.} MAE(Corr_i,\ k_i) \qquad (17)$$

The values of $Corr^*$ and k^* are potentially the most effective in reflecting the characteristics of the distribution of the presence of substantive-content message properties in the information background. The developed algorithm is presented in Figure 2.

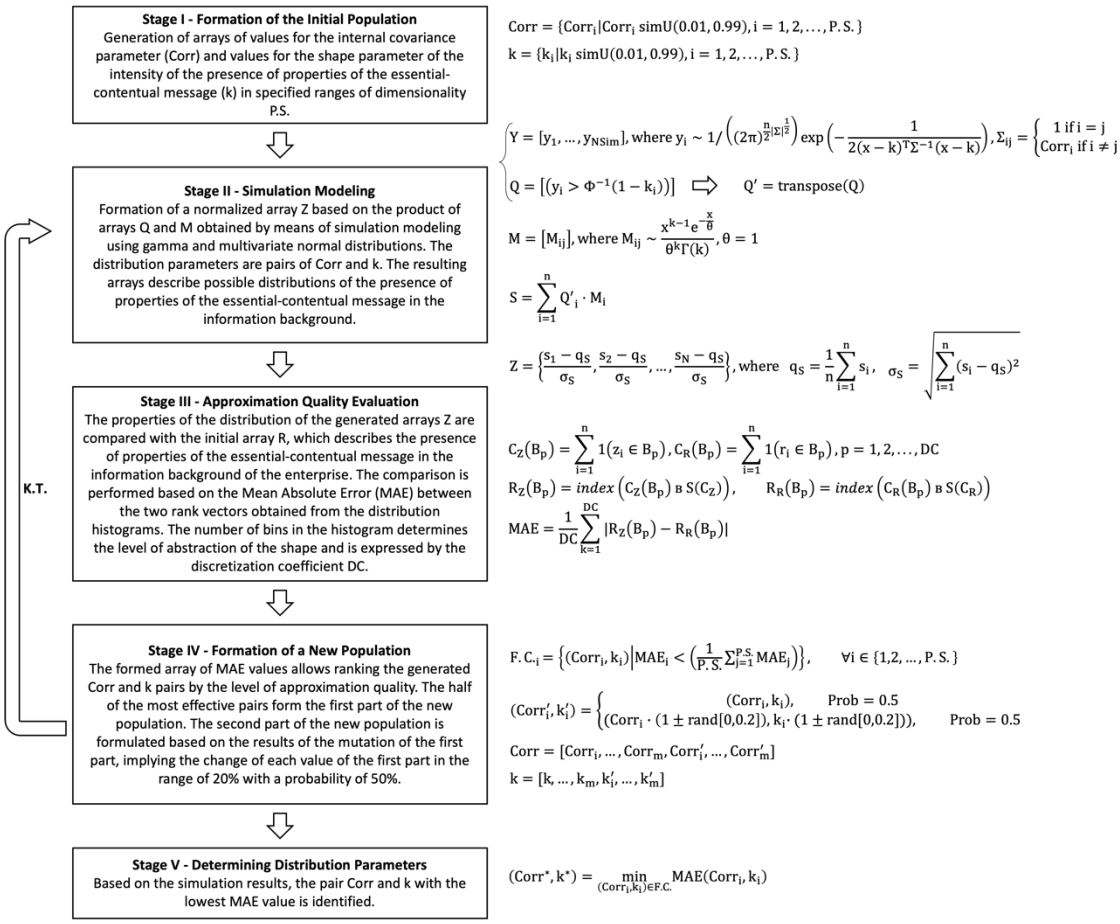

Figure 2. Algorithm for modeling the distributions of the presence of substantive-content message properties in the information background

The analytical data obtained as a result of the reconciliation of this algorithm will allow us to draw conclusions about the intensity and specificity of a particular property of the substantive-content message in the information background.

The model can be validated using real-world datasets from domains such as social media or political communication, where multimodal information patterns are prominent (Biggers et al., 2023; Lytras et al., 2020). For instance, analyzing the distribution of substantive-content message properties in datasets derived from online interactions or advertising campaigns could showcase the model's capability in real-world scenarios (Boerman et al., 2021). Comparing model results with real data shows how accurate it is.

A comparative review of similar approaches, such as those utilizing Latent Dirichlet Allocation (LDA) (Constantiou & Kallinikos, 2015) or clustering algorithms like K-Means and DBSCAN (Cawsey & Rowley, 2016; Lytras et al., 2020), will highlight the improvements introduced by integrating Gamma and multivariate normal distributions optimized via genetic algorithms. These enhancements, including better handling of multimodal distributions and complex thematic interdependencies, can position the proposed method as a significant advancement over existing methodologies (Goldberg, 2022).

The proposed model is particularly suited for analyzing datasets from social media and political news, where the information background often exhibits multimodal characteristics. Social media platforms generate vast amounts of user-generated content, including posts, comments, and interactions, which can be used to detect dominant themes, measure emotional intensity, and identify patterns in user engagement (Biggers et al.,

2023). Similarly, datasets from political news, which often reflect highly polarized and event-driven dynamics, provide an opportunity to validate the model's ability to capture the intensity and frequency of recurring themes (Boerman et al., 2021).

The methodology proposed in this study offers several distinct advantages compared to existing alternatives for analyzing the presence of substantive-content message properties in information backgrounds. Unlike traditional tools such as Latent Dirichlet Allocation (LDA) or K-Means clustering, which often treat information properties in isolation or assume linear separability in data, the developed model integrates both Gamma distribution for intensity and multivariate normal distribution for frequency. This dual-probabilistic framework enables a more nuanced representation of multimodal distributions, capturing both the depth (intensity) and recurrence (frequency) of information properties.

Moreover, existing approaches like LDA are limited in their ability to handle the interdependencies between properties and often fail to account for contextual relevance. In contrast, the proposed model explicitly incorporates interrelationships through the covariance matrix in the multivariate normal distribution, allowing it to analyze correlations and co-occurrence patterns more effectively. Additionally, while methods like DBSCAN and Mean-Shift clustering excel in identifying data density and clusters, they do not account for the semantic or thematic content of the information. The developed methodology bridges this gap by integrating genetic algorithms for parameter optimization, ensuring adaptability to complex, multidimensional data structures and improving model accuracy.

Furthermore, the model's robustness to noise, outliers, and incomplete data provides a significant advantage over alternative tools that often rely on clean and structured datasets. This feature ensures reliable performance in real-world scenarios, such as social media analysis or political communication, where data is inherently noisy and incomplete.

For instance, the model can be applied to social media data to identify key topics and assess their saturation and recurrence in different communities or demographics. In political news analysis, the methodology can track shifts in narrative focus or sentiment over time, highlighting the correlation between certain topics and audience reactions. By leveraging Gamma distribution for intensity and multivariate normal distribution for frequency, the model can effectively represent the diverse properties of information dynamics in these contexts.

The proposed methodology is designed to address common challenges in information environments, including noise, outliers, and incomplete data, ensuring robustness and reliability. By employing the Gamma distribution for intensity modeling, the approach inherently smooths over random fluctuations in data, minimizing the impact of noise through its shape and scale parameters. Outliers are effectively managed through the multivariate normal distribution, where the covariance matrix identifies and reduces the influence of anomalous data points by accounting for established correlations. Similarly, the shape parameter (k) in the Gamma distribution adjusts for dispersion, mitigating the effect of extreme intensity values. To handle incomplete data, the methodology incorporates imputation techniques during preprocessing, maintaining the integrity of the data array for analysis. Additionally, the genetic algorithm's iterative optimization process is resilient to missing values, as it seeks global optima across the parameter space, compensating for data gaps. These features collectively enable the model to deliver accurate and meaningful insights even in noisy, outlier-prone, or incomplete data environments.

This study has yielded the following results:

1. The algorithm (Figure 1) effectively captures the multimodal distribution of substantive-content message properties by combining intensity and frequency parameters.
2. The approach models the presence of properties as peaks of varying significance and recurrence, providing a nuanced understanding of the information landscape.
3. The intensity parameter, modeled with the Gamma distribution, quantifies the depth and saturation of substantive-content messages. This highlights how specific messages stand out in terms of their emotional or thematic richness, reflecting the strength and impact of individual topics (Figure 1, "Intensity Modeling").
4. The frequency parameter, derived from the multivariate normal distribution, accounts for the co-occurrence patterns and correlations between different topics. This enables the detection of relationships between topics and the frequency with which they appear together, revealing the interconnected nature of information signals (Figure 1, "Frequency Modeling").
5. A genetic algorithm optimizes the parameters of intensity and frequency distributions to achieve the best fit to the observed data. By iteratively improving the model, the genetic algorithm identifies global optima, ensuring robustness and adaptability to complex, multidimensional datasets (Figure 1, "Genetic Algorithm Optimization").
6. The combination of intensity and frequency parameters enables the identification of dominant themes and their interrelationships. The algorithm (Figure 1, "Generate Analytical Insights") supports in-depth analysis of the structure and dynamics of the information background, paving the way for more effective decision-making.
7. The methodology is particularly suited for applications in social media analysis, political communication, and marketing. By forecasting trends and understanding the interplay of key themes, the algorithm offers valuable insights for strategic planning and operational improvements (Figure 1, "Practical Applications").

4. Conclusions

This study has presented a novel methodology for modeling the distribution of substantive-content message properties in the information background. The methodology leverages the concept of multimodality in the distribution of these properties, which is characterized by peaks of varying intensity and frequency. The intensity of the presence of a property is defined as the depth or saturation of the information signal associated with it, while the frequency represents the number of times it appears in the information background.

To model these aspects, we propose a combined approach using the Gamma and multivariate normal distributions. The Gamma distribution effectively captures the intensity of presence, allowing us to estimate the expected number of occurrences and the time scale of these events. Meanwhile, the multivariate normal distribution accounts for the frequency of presence, capturing the complex relationships between different substantive-content message properties.

A key advantage of this approach is its ability to model the interplay between intensity and frequency, thus providing a more comprehensive understanding of the information background. To effectively identify the parameters of these distributions, we employed a genetic algorithm approach, enabling us to explore the multidimensional parameter space and identify global optima.

The proposed methodology offers several potential applications, including:

• Identifying the dominant themes and narratives within a specific information space.

• Assessing the intensity and frequency of specific topics, events, or trends.

• Understanding the interrelationships between various substantive-content message properties.

• Developing predictive models for the evolution of information dynamics.

By incorporating both intensity and frequency as key parameters, this study provides a more refined and nuanced approach to analyzing the information background. This approach is especially relevant in today's information-saturated environment, where understanding the complexities of information distribution is crucial for effective decision-making.

Future research can further enhance this methodology by:

• Exploring additional distributions to model different types of substantive-content message properties.

• Developing more sophisticated approaches for handling complex interdependencies between properties.

• Investigating the potential applications of this methodology in various domains, such as social media analysis, political communication, and marketing.

Overall, the proposed methodology represents a valuable tool for understanding the complex interplay of substantive-content message properties within the information background. It offers a robust framework for analyzing and quantifying information dynamics, paving the way for deeper insights and more informed decision-making in various domains.

Acknowledgments

The research is financed as part of the project “Development of a methodology for instrumental base formation for analysis and modeling of the spatial socio-economic development of systems based on internal reserves in the context of digitalization” (FSEG-2023-0008).

Conflict of Interest

The authors declare no conflicts of interest.

References

Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6), pp. 3911. DOI: https://doi.org/10.3390/app13063911

Anisiforov A. B., Dubgorn A. S. (2017). Organization of enterprise architecture information monitoring. *Proceedings of the 29th International Business Information Management Association Conference-Education Excellence and Innovation Management through Vision 2020: From Regional Development Sustainability to Global Economic Growth*, pp. 2920-2930. DOI: https://doi.org/10.1051/e3sconf/201911002051

Bashir, H., Ojiako, U., Marshall, A., Chipulu, M., & Yousif, A. A. (2022). The analysis of information flow interdependencies within projects. Production Planning & Control, 33(1), 20-36. DOI: https://doi.org/10.1080/09537287.2020.1821115

Berawi, M.A., Sari, M., Salsabila, A.A., Susantono, B., Woodhead, R., 2022. Utilizing Building Information Modelling in the Tax Assessment Process of Apartment Buildings. International Journal of Technology. Volume 13(7), pp. 1515-1526. DOI : https://doi.org/10.14716/ijtech.v13i7.6188

Beth, E. W. (Ed.). (2012). Formal methods: an introduction to symbolic logic and to the study of effective operations in arithmetic and logic *(Vol. 4). Springer Science & Business Media*.

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: an empirical investigation. *MIS quarterly*, pp. 169-196. DOI: 10.2307/3250983.

Biggers, F. B., Mohanty, S. D., & Manda, P. (2023). A deep semantic matching approach for identifying relevant messages for social media analysis. *Scientific Reports*, 13(1), pp. 12005. DOI: https://doi.org/10.1038/s41598-023-38761-y

Boerman, S. C., Kruikemeier, S., & Bol, N. (2021). When is personalized advertising crossing personal boundaries? How type of information, data sharing, and personalized pricing influence consumer perceptions of personalized advertising. *Computers in Human Behavior Reports*, 4, pp. 100144. DOI: https://doi.org/10.1016/j.chbr.2021.100144

Bruce, N. I., Murthi, B. P. S., & Rao, R. C. (2017). A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement. Journal of marketing research, 54(2), 202-218. DOI: 10.1509/jmr.14.0117

Cappella, J. N., & Li, Y. (2023). Principles of effective message design: A review and model of content and format features. Asian Communication Research, 20(3), 147-174. DOI: 10.20879/acr.2023.20.023

Cawsey, T., & Rowley, J. (2016). Social media brand building strategies in B2B companies. *Marketing Intelligence & Planning*, 34(6), pp. 754-776. DOI: https://doi.org/10.1108/MIP-04-2015-0079

Clark, D., Hunt, S., & Malacaria, P. (2007). A static analysis for quantifying information flow in a simple imperative language. Journal of Computer Security, 15(3), 321-371. DOI: 10.3233/JCS-2007-15302

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. *Journal of Information Technology*, 30(1), 44-57. DOI: 10.1057/jit.2014.17.

Damanpour, F. (2018). Organizational innovation: A meta-analysis of effects of determinants and moderators. *In Organizational innovation*. pp. 127-162. Routledge.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6), pp. 254-262. DOI: https://doi.org/10.1016/j.tics.2011.04.003

Deutsch, K. W. (1951). Mechanism, organism, and society: Some models in natural and social science. *Philosophy of Science*, 18(3), pp. 230-252.

Dretske, F. (1981). Knowledge and the Flow of Information.

Drucker P. F. (1990) Butterworth Heinemann. *Managing for the future - and beyond.* T . 281.

Eremina, I., Yudin, A., Tarabukina, T., Oblizov, A., 2022. The Use of Digital Technologies to Improve the Information Support of Agricultural Enterprises. International Journal of Technology. Volume 13(7), pp. 1393-1402. DOI : https://doi.org/10.14716/ijtech.v13i7.6184

Gliboff, S. (1999). Gregor Mendel and the laws of evolution. *History of Science*, 37(2), pp. 217-235.

Goldberg, Y. (2022). Neural Network Methods for Natural Language Processing. DOI: https://doi.org/10.1007/978-3-031-02165-7

Hitt M. A., Ireland R. D., Hoskisson R. E. (2019) Strategic management: Concepts and cases: Competitiveness and globalization. Cengage Learning.

Hsieh, Y. C., & Chen, K. H. (2011). How different information types affect viewer's attention on internet advertising. *Computers in human Behavior*, 27(2), pp. 935-945. DOI: 10.1016/j.chb.2010.11.019

Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: *A literature review using text mining. International Journal of Information Management Data Insights*, 1(2), pp. 100017. DOI: 10.1016/j.jjimei.2021.100017

Li, Y., Chang, Y., & Liang, Z. (2022). Attracting more meaningful interactions: The impact of question and product types on comments on social media advertisings. *Journal of Business Research*, 150, pp. 89-101. DOI: 10.1016/j.jbusres.2022.05.085

Lim, S., & Schmälzle, R. (2023). Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Frontiers in Communication, 8, 1129082. DOI: https://doi.org/10.3389/fcomm.2023.1129082

Lytras, M., Visvizi, A., Zhang, X., & Aljohani, N. R. (2020). Cognitive computing, Big Data Analytics and data driven industrial marketing. *Industrial Marketing Management*, 90, pp. 663-666. DOI: 10.1016/j.indmarman.2020.03.024

Mueller, M., & Grindal, K. (2019). Data flows and the digital economy: information as a mobile factor of production. *Digital Policy, Regulation and Governance*, 21(1), pp. 71-87. DOI: 10.1108/DPRG-08-2018-0044

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), pp. 544-551. DOI: 10.1136/amiajnl-2011-000464

Nugraha, E., Sari, R.M., Sutarman, Yunan, A., Kurniawan, A., 2022. The Effect of Information Technology, Competence, and Commitment to Service Quality and Implication on Customer Satisfaction. International Journal of Technology. Volume 13(4), pp. 827-836. DOI : https://doi.org/10.14716/ijtech.v13i4.3809

Öberg, C. (2023). Neuroscience in business-to-business marketing research: A literature review, co-citation analysis and research agenda. *Industrial Marketing Management*, 113, pp. 168-179. DOI: 10.1016/j.indmarman.2023.06.004

Onykiy, B., Antonov, E., Artamonov, A., & Tretyakov, E. (2020). Information analysis support for decision-making in scientific and technological development. *International Journal of Technology*, 11(6), pp. 1125-1135. DOI : https://doi.org/10.14716/ijtech.v11i6.4465

Pribram, K. H. (1960). A review of theory in physiological psychology. *Annual review of psychology*, 11(1), pp. 1-40.

Qiu, Q., Hao, Z., & Jiang, L. (2022). Strategic information flow under the influence of industry structure. *European Journal of Operational Research*, 298(3), pp. 1175-1191. DOI: 10.1016/j.ejor.2021.08.041

Rodionov, D., Gracheva, A., Konnikov, E., Konnikova, O., & Kryzhko, D. (2022). Analyzing the systemic impact of information technology development dynamics on labor market transformation. *Int. J. Technol*, 13(7), pp. 1548-1557. DOI: https://doi.org/10.14716/ijtech.v13i7.6204

Rodionov, D., Kryzhko, D., Tenishev, T., Uimanov, V., Abdulmanova, A., Kvikviniia, A., ... & Konnikov, E. (2022). Methodology for assessing the digital image of an enterprise with its industry specifics. *Algorithms*, 15(6), pp. 177. DOI: https://doi.org/10.3390/a15060177

Rodionov, D., Zaytsev, A., Konnikov, E., Dmitriev, N., & Dubolazova, Y. (2021). Modeling changes in the enterprise information capital in the digital economy. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(3), pp. 166. DOI: https://doi.org/10.3390/joitmc7030166

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), pp. 379-423.

Stapleton, G., Howse, J., & Rodgers, P. (2010). A graph theoretic approach to general Euler diagram drawing. *Theoretical Computer Science*, 411(1), pp. 91-112. DOI: https://doi.org/10.1016/j.tcs.2009.09.005

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society, Series 2*, 45, pp. 161-228.

Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1), pp. 106. DOI: https://doi.org/10.1038/s41597-023-01995-6

Weaver, W. (2017). The mathematics of communication. *In Communication theory* (pp. 27-38). Routledge.

Weinlich P., Semerádová T. (2022). Emotional, cognitive and conative response to influencer marketing. *New Techno Humanities*. 2(1), pp. 59-69.

Zaytsev A. et al. Comparative analysis of results on application of methods of intellectual capital valuation //*International scientific conference «Digital transformation on manufacturing, infrastructure and service»(DTMIS 2019).* St. Petersburg. – 2019. DOI: 10.1088/1757-899X/940/1/012025

Appendix - Methodology for Modeling the Properties of the Substantive-Content Message Encoded in the Form of Symbolic Data Constructs