Contextual-formal concept of creating explainable models for corpus linguistics tasks

Viacheslav Kovtun[1,†], Victoria Vysotska[2,*,†] and Oksana Kovtun[3,†]

[1] *Vinnytsia National Technical University, Khmelnytske shose, 95, Vinnytsia, 21021, Ukraine*

[2] *Information Systems and Networks Department, Lviv Polytechnic National University, Stepan Bandera Street, 12, Lviv, 79013, Ukraine*

[3] *Vasyl' Stus Donetsk National University, 600-richchya Str., 21, Vinnytsia, 21000, Ukraine*

Abstract

The article proposes an explainable model based on a δ-relaxed formal structure, which provides a transparent interpretation of classification decisions through the ontological structure of δ-concepts. The theoretical basis of the model is a δ-modified formal conceptual analysis with support for partially implemented features and fuzzy relationships between objects and attributes. An aggregated interest function is introduced, focused on optimising the semantic consistency of δ-concepts. The model is implemented as a classifier with an explanatory layer based on δ-concepts, tested on a corpus of Ukrainian utterances in the task of automatic recognition of pragmatic types. The model demonstrated high efficiency: F1-score – 0.84, average Lift – 1.23, Δ-Stability – 0.77, label entropy – 0.50. Statistical analysis showed a significant advantage of the δ-model in terms of Lift ($p = 0.049$) compared to CBM, which confirms more effective detection of informative concepts without loss of accuracy. The practical significance of the study lies in the creation of interpretable models for chatbots, educational systems, and legal analysis.

Keywords

δ-relaxed formal model, explainable architecture, concept lattice, interest function, pragmatic sentence classification, semantic features, interpretability in NLP, statistical significance

1. Introduction

The interpretability of AI decisions is a critical condition for the implementation of language technologies in areas of increased social and legal responsibility [1-3]. In such areas as e-justice, medical consulting, education, and moderation of public discussions, the classification decision must be not only accurate, but also transparent for the end user – a lawyer, doctor, teacher, or moderator. For example, in the analysis of court decisions to identify language patterns of bias, it is necessary to explain which language constructs were the basis for the classification. In medical chatbots, it is essential not only to provide an answer but also to argue why a particular directive is interpreted as a request, and not as an independent decision. In educational platforms, the classification of the type of student statement (question, statement, doubt) affects the adaptation of academic content, which requires formal justification of the model's actions.

Despite technical progress in the field of transformative architectures, most language models do not provide the ability to trace which features (explicit, implicit, or partial) influenced the classification [4-6]. Natural language is not reduced to complete binary relationships between words and meanings: it is common to encounter cases of incomplete or implicit implementation of semantic features, as well as context-dependent variations. It gives rise to an open scientific problem: how to

MoDaST 2025: Modern Data Science Technologies Doctoral Consortium, June, 15, 2025, Lviv, Ukraine

[1*] Corresponding author.

[†] These authors contributed equally.

kovtun_v_v@vntu.edu.ua (V. Kovtun); victoria.a.vysotska@lpnu.ua (V. Vysotska); o.kovtun@donnu.edu.ua (O. Kovtun)

0000-0002-7624-7072 (V. Kovtun); 0000-0001-6417-3689 (V. Vysotska); 0000-0002-9139-8987 (O. Kovtun)

build an explainable architecture that combines flexibility in feature representations with formal transparency of decisions, ensuring interpretation even in cases of unclear implementation of linguistic information. There is a need to create an explainable model that will provide an explanation of classification decisions in a formalised form. The focus of the study is the classification of Ukrainian utterances by pragmatic types (statements, directives, questions, expressives), within which the explanation of communicative intention requires not only identification but also structural justification.

The explainability of classification decisions in natural language processing is critically important in the context of pragmatic utterance analysis. Existing classification architectures often demonstrate high accuracy, but remain opaque about which features influenced the assignment of an utterance to a particular pragmatic type. It is especially complicated in cases of partial, implicit, or variable implementation of linguistic features. We will conduct a systematic critical analysis of modern approaches to explaining the decisions of NLP models.

Modern large language models (GPT-4, Claude, Gemini) implement explainability through textual generation of reasoning, driven by special instructions (prompt-based explanation). The most common methods are Chain-of-Thought prompting [7], Self-Generated Rationales [8], and Instruction Fine-Tuning [9]. They are used to explain classification decisions in the form of sequential logical phrases, in particular when defining pragmatic types of statements (questions, statements, directives). Although such explanations are intuitive, they are uncontrolled, simulative in nature and are not based on the deterministic logic of the model. In most cases, they depend on patterns learned during training, rather than on interpreted internal representations. Variable results when reformulating a query, instability of rationalisation, and lack of reproducible formalised structure limit their application in tasks where transparent and controlled interpretation is required.

Post-facto interpretation of deep learning results in NLP is most often implemented by LIME [10], SHAP [11], Integrated Gradients [12] and Attention Rollout [13] methods. These approaches provide a local assessment of the importance of features or tokens, based on gradient influence or approximation of the original model by linear interpreters. Their advantage is fast integration with non-interpretable models, such as BERT, RoBERTa, and DistilBERT, without modifying the architecture. For example, in classifying statements as directives or questions, these methods can determine which tokens influenced the decision. However, the explanations are unstable when paraphrasing, do not have a common logical structure, and do not guarantee compliance with the syntactic-semantic implementation of the features. In addition, such methods do not allow the detection of implicitly implemented or partial features, which is critical in pragmatic analysis.

Formal rule-based approaches are implemented in the form of decision trees, IF–THEN patterns, and rule-based frameworks such as SlugNERDS [14] and RuleBERT [15]. In such models, each classification decision is accompanied by a precise sequence of logical conditions, which allows for full decision tracing. The advantage is high interpretability and control over the classification logic, which will enable you to accurately determine which linguistic features, for example, the presence of an imperative mood or an interrogative pronoun, became the basis for attributing a sentence to a specific pragmatic type. At the same time, such models are rigid to variable language constructions and are unable to adequately process statements with implicit or partial implementation of features. The scalability of rule-based systems to open-domain environments or unstructured corpora is limited.

Explanation methods using graph representations are based on the construction of semantic graphs or Graph Attention Networks (GAT) [16], which allow interpreting model solutions through the relationships between concepts. Knowledge graph-based reasoning (e.g., KG-XAI [17]) and path-based explanations are also used in NLP tasks to form reasoning trajectories. The advantage is the ability to visualise complex dependencies, integration with external ontologies (ConceptNet, FrameNet [18]), and the use of relationships between concepts as a basis for explaining classification. However, the application of such approaches to sentence analysis in natural language is complicated: the construction of graphs requires a clear knowledge structure, while statements can have implicit

or fuzzy semantics. In addition, there is no direct connection between the vertices of the graph and the grammatical structure of the sentence, which makes it impossible to fully trace solutions.

The Formal Concept Analysis (FCA) method [19] is used to construct ontological structures based on binary relations between objects and features. In the field of NLP, FCA is used for thematic classification, generalisation of semantic constructs, and formation of concept lattices [20, 21]. The main advantage is the ability to construct an interpreted concept lattice that describes sets of objects with the same features. It allows classification logic to be formalised in the form of explainable rules, as well as generalised statements by semantically similar features. However, classical FCA works with a rigid binary matrix that assumes full implementation of features in each object. In cases of partial, variable, or implicit implementation of features, which is typical for pragmatic sentences, such a model loses relevance or requires excessive discretisation, which leads to the loss of meaningful information.

Therefore, in the context of analysing the advantages and disadvantages of classical approaches to the explainable classification of pragmatic statements, the use of δ-relaxed models with built-in interest functions is promising. Post-factum methods and generative LLM tools provide only local simulated explanation, rule-based and FCA approaches are rigid with respect to partially realised features, and graph-based solutions require external ontologies and do not formalise the logic of reasoning. Against this background, δ-relaxed formalisation allows combining structural interpretability, fuzziness, and explainability, which makes it a relevant basis for building explainable architectures in the tasks of semantic-pragmatic analysis of natural language texts.

The object of the research is the process of explainable classification of pragmatic statements in natural language with partially implemented semantic features.

The subject of the research is the theoretical foundations, formal models and methods of constructing an interpreted classification of pragmatic statements in natural language, taking into account the partial implementation of semantic features. It includes the analysis of modern explainable approaches (prompt-based, post-hoc, rule-based, graph-based, FCA) and the development of a δ-relaxed conceptual model with interest functions, which explains decisions based on the semantic and grammatical characteristics of sentences.

The research aims to improve the interpretability and flexibility of the process of classifying pragmatic statements by developing a δ-relaxed model capable of formally reflecting the partial implementation of semantic features in a natural language corpus based on conceptual structures and interest functions.

The article is structured as follows: Section 2 presents the theoretical basis of the study: a formalisation of the δ-relaxed model of explanatory classification of pragmatic statements is carried out, which takes into account the partial implementation of semantic features in the natural language environment. A generalised incident relation is proposed, formal definitions of δ-concepts, interest functions (in particular, target entropy, Δ-stability, Lift), as well as analytical assessments that provide a ranking of generalisations according to their explanatory potential, are proposed. Section 3 presents the results of the experimental study: a corpus of pragmatic statements is described, the construction of δ-relaxed lattices is implemented, interest analysis is performed, and examples of interpretations of classification solutions with numerical quality assessments (measures of generalisation, information gain, stability, etc.) are demonstrated. Section 4 summarises the conclusions of the study: the scientific novelty is highlighted, the effectiveness and practical value of the proposed model are confirmed, its limitations are identified, and promising directions for further development are outlined.

2. Models and Methods

Classical FCA is a powerful tool for detecting hidden categories in tabular structures. However, its application in computational linguistics is limited by a rigid binary logic: either a feature is present or it is not. In contrast, in natural language, features can be partially realised, with vagueness or instability of detection, for example, in the case of co-occurrence structures (statistically significant

co-occurrences of words, such as "sharp criticism"), synonymous variations, or pragmatic shifts. Therefore, there is a need for a δ-relaxed model, where the correspondence between an utterance and a feature allows for a gradual degree of membership, controlled by the parameter δ. In corpus linguistics, many features do not have the rigid binary nature predicted by classical FCA. Categories such as modality, pragmatic shift, co-occurrence relevance, or synonymy are not detected with absolute certainty, but with a certain degree of probability or fuzziness. It necessitates a δ-relaxed model that allows the use of the degree of membership of a feature to an utterance within the interval [0,1], which allows for accurate modelling of fuzzy or partial manifestations of linguistic properties.

In corpus linguistics, a set of objects S is considered a collection of tokenised sentences or contexts, utterances or fragments of discourse. The set of features F includes morphological categories, part-of-speech tags, syntactic relations, semantic labels (e.g. modality, movement, agentivity) or co-occurrence patterns. Formal context is defined as

$$C = (S, F, J), J \subseteq S \times F, \quad (1)$$

where $(s, f) \in J$ if and only if the feature f is inherent in the statement s with high confidence. For fuzzy data, a δ-relaxed version (1) is introduced: $C_\delta = (S, F, J_\delta)$, $J_\delta \subseteq S \times F \times [0,1]$, where $\mu_\delta(s, f) \in [0,1]$ is the degree of correspondence of the feature f to the sentence s. In general, the context J or J_δ (in the case of the δ-model) represents the relationship between a set of sentences (statements) and a set of features (lexical, morphological, semantic, syntactic, etc.). This structure corresponds to a sparse table or matrix, where the values can be both binary (0/1) and gradational (fuzzy), which will be reflected in the examples in Section 3.

Closure operators allow us to associate sets of sentences with corresponding sets of features: $K^{'} = \{f \in F | \forall s \in K: (s, f) \in J\}$, $K \subseteq S$, $L^{'} = \{s \in S | \forall f \in L: (s, f) \in J\}$, $L \subseteq F$. In the δ-context, fuzzy closure operators are introduced: $K'_\delta = \{f \in F | \mu_\delta(s, f) \geq \delta, \forall s \in K\}$, $L^{'}_\delta = \{s \in S | \mu_\delta(s, f) \geq \delta, \forall f \in L\}$.

On the basis of classical FCA (i.e. over C), the formal concept is interpreted as a pair (K, L), which satisfies the two-way closure condition:

$$K^{'} = K, L^{'} = L, K \subseteq S, L \subseteq F. \quad (2)$$

It means that all sentences with K have standard features of L, and vice versa – L characterises only these sentences. In the following, unless otherwise stated, all pairs (K, L) are interpreted relative to a fixed context C or C_δ. In δ-cases, condition (2) may have a fuzzy implementation.

The generalisation relation can order the concepts:

$$(K_1, L_1) <= (K_2, L_2) \Leftrightarrow K_1 \subseteq K_2, L_2 \subseteq L_1. \quad (3)$$

Thus, concepts with smaller scope and more specialised features are concretisations of more general ones. As the corpus size or number of features increases, a combinatorial explosion of concepts is observed. Therefore, there is a need for cognitively oriented filtering methods based on interest indices, which we will consider later.

The coverage ratio in a lattice is given by:

$$(K_1, L_1) \prec (K_2, L_2) \Leftrightarrow K_1 \subset K_2, L_2 \subset L_1, \nexists (K_3, L_3): K_1 \subset K_3 \subset K_2. \quad (4)$$

In the neural network interpretation, expression (4) corresponds to a direct connection between nodes (neurons) that implement the transition from a general pattern to a more specific one. In practical cases, concepts that have only a partial overlap of the volumes $K_1 \cap K_2 \neq \emptyset$, but are not in a covering relation, are often encountered. To ensure the coherence of the δ-lattice, δ-bridges are introduced - auxiliary concepts that connect structurally close, but formally uncovered pairs:

$$\left(K_{1,2}, L_{1,2}\right) := \left(K_1 \cap K_2, (K_1 \cap K_2)^{'}_\delta\right), \quad (5)$$

where $L_{1,2}$ – is formed by the δ-closure of the intersection $K_{1,2}$. However, the reverse closure may not be fulfilled. Such pairs may not satisfy the strict conditions of a formal concept, but they play the role of connecting units in δ-structures. δ-bridges of the form (5) play the role of cognitively relevant transitional concepts that connect concepts with a partial but significant intersection of volumes. Although such pairs do not satisfy the conditions of a formal concept in the classical sense, they ensure the coherence of the δ-lattice and the continuity of the semantic structure between concepts. In the context of explainable architectures, δ-bridges play the role of buffer nodes that contribute to the understandable interpretation of the internal layers of the model. Their effectiveness will be analysed empirically in the next section.

In the tasks of building neural networks based on lattices of formal concepts, filtering concepts is of particular importance, which will subsequently form the structure of hidden layers. Due to the exponential growth of the number of concepts with an increase in the size of the input space, there is a need to select the most relevant ones. This task is solved through interest indices - numerical functions that rank formal concepts by their cognitive, statistical or logical significance. Let us present the interpretation of interest indices, the most relevant for the δ-model of formal analysis of linguistic concepts, and the most common indices.

Basic Level Index has a cognitive motivation, a concept must have high internal coherence, be more coherent than its superconcepts and not much less coherent than its subconcepts:

$$\beta(K,L) = t\big(\kappa_1(K,L), \kappa_2(K,L), \kappa_3(K,L)\big), \tag{6}$$

where $t(\cdot)$ is the aggregating t-norm (for example, the product), $\kappa_1(K,L)$ is the coherence of the concept itself, $\kappa_2(K,L)$ is the comparison with superconcepts, $\kappa_3(K,L)$ is the comparison with subconcepts. Coherence is calculated as:

$$\gamma_{\varnothing}(K,L) = \frac{1}{\frac{|K|(|K|-1)}{2}} \sum_{\substack{\{e_1,e_2\}\subset K,\\ e_1 \neq e_2}} \mathrm{sim}(e_1,e_2), \tag{7}$$

where $L_e = \{f \in F | (e,f) \in J\}$ is the set of features belonging to object e, and $\mathrm{sim}(e_1,e_2)$ is the similarity between two objects is based on their features. The metrics used are:

$$\mathrm{sim}_{SMC}(L_1,L_2) = \frac{(|L_1 \cap L_2| + |X - (L_1 \cup L_2)|)}{|X|}, \tag{8}$$

$$\mathrm{sim}_{J}(L_1,L_2) = \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}, \tag{9}$$

where $X \subseteq F$ is a subset of features by which similarity is calculated.

In the context of (7), the functions $\kappa_{\{1,2,3\}}(K,L)$ mentioned in (6) are defined as:

$$\kappa_1(K,L) = \gamma_{\varnothing}(K,L),\ \kappa_2(K,L) = 1 - \frac{1}{|UN(K,L)|} \sum_{y \in UN(K,L)} \frac{\gamma_{\emptyset}(y)}{\gamma_{\emptyset}(K,L)}, \tag{10}$$

$$\kappa_3(K,L) = 1 - \frac{1}{|LN(K,L)|} \sum_{y \in LN(K,L)} \frac{\gamma_{\varnothing}(K,L)}{\gamma_{\varnothing}(y)},$$

where $UN(K,L)$, $LN(K,L)$ are the sets of immediate super- and subsumed pairs (K,L) in the complete lattice of formal concepts constructed over the context C (see (3)). We will denote the realisations of index (6) as β_S (based on SMC) and β_J (based on Jacquard).

The Target Entropy index is used in classification problems where each object has a target label. It is defined as the variance or entropy of classes among objects of a concept:

$$H(K,L) = D(tl(s), s \in K), \tag{11}$$

where $D(\cdot)$ is the variance, and $tl(s) \in C$ is the target label of an object s belonging to a set of classes C (e.g., tonality, topic, pragmatic type). The lower the entropy, the more consistent the concept is with a particular class.

The classical stability of a concept characterises the stability of its features to variations of a set of objects. The computational complexity of the process of assessing stability in practice uses the approximate Δ-Stability index:

$$\Delta(K,L) = \min_{p<q}\left(|K| - \left|K_p\right|\right), \tag{12}$$

where $q = (K,L)$, p is any subconcept of the pair q such that $K_p \in K$. The index Δ shows how "distant" the concept is from its closest subconcepts and, accordingly, how structurally stable it is.

The Lift index is used in association analysis to identify non-obvious but statistically informative associations between features.:

$$\Lambda(K,L) = \frac{\prod_{l=L} Pr(l)}{Pr(L)}, \tag{13}$$

where $Pr(l) = \frac{|l'|}{|S|}$ is the probability of occurrence of feature $l \in L$, $Pr(L) = \frac{|L'|}{|S|}$ is the probability of the joint occurrence of all features L. If $\Lambda(K,L) > 1$, the features from the set L are positively correlated. To avoid numerical instability, the logarithmic version (13) is often used:

$$log\,\Lambda\,(K,L) = \sum_{l\in L} logPr(l) - logPr(L). \tag{14}$$

Among the above indices, the most potentially suitable for explainable architectures are β_J (reflecting internal coherence) and Δ-Stability (providing structural separation). Their combination allows for the formation of hidden layers of the neural network as interpreted concepts, consistent in features, and isolated in a lattice topology.

In explainable AI tasks for natural language processing, a critical stage is the detection of interpreted concepts that form the structure of the hidden layers of the neural network. Unlike tabular data, language corpora are characterised by complex semantics, uneven feature structure and high variability, which makes it impossible to use classical association rules or rigid ontologies. Therefore, the goal is to construct a set of δ-concepts $C_\delta\{(K_i,B_i)\}_{i=1}^{c} \subseteq I_\delta$, which, taking into account the relaxation of the membership relation, cover S and correspond to high values of the interestingness indices (expressions (6)–(f14)). Here c is a natural number that denotes the number of selected δ-concepts (concepts) in the set C_δ, I_δ is a δ-lattice of formal concepts, that is, the set of all pairs (K,L) that satisfy the condition of partial correspondence between objects and features with an accuracy of at least δ: $I_\delta = \left\{(K,L)\,\middle|\, K \subseteq S, L \in F, \forall s \in K: \frac{|\{f\in L|(s,f)\in J\}|}{|L|} \geq \delta\right\}$. Each concept in the structure of the set of δ-concepts is a potential node in the neural network model.

In the sought-after δ-context $C_\delta = (S,F,J_\delta)$, the membership relation $J_\delta \subseteq S \times F$ is defined as a partial correspondence: an object $s \in S$ satisfies the feature set $L \subseteq F$ if:

$$\frac{|\{f\in L|(s,f)\in J\}|}{|L|} \geq \delta. (\tag{15}$$

Inequality (15) must hold for each $s \in K$, i.e., the features L must be characteristic of all objects of the set K with an accuracy of at least δ.

The construction of the lattice is preceded by a procedure for preprocessing the corpus, which includes:

– tokenisation, lemmatisation, PoS-markup;

– generation of features F, including grammatical (gender, number, tense), syntactic (subject, predicate, object) and semantic (action, place, evaluation) markers;

– construction of C_δ taking into account the given level of fuzziness δ.

Based on C_δ, a δ-lattice of concepts I_δ is constructed. In this case, the closures $K' \approx L$, $L' \approx K$ are understood as partial, taking into account the fuzzy relation J_δ. It allows maintaining consistency with the basic theory of formal concepts when transitioning to a fuzzy model.

Each concept $(K,L) \in I_\delta$ is a candidate for the role of a node of the explainable architecture. The choice of a specific index or its combination depends on the task:

– Basic Level Index, in particular its implementations β_S (based on Simpson's metric) and β_J (based on Jaccard's metric). It is recommended for detecting concepts with high internal coherence that are well-matched by features. The index is defined by formula (6), coherence components (7)–(10), including the corresponding similarity metrics (8), (9), and aggregation is performed via t-norm.

– Target Entropy is used when objects have target labels (classes), for example, in classification or clustering tasks. For numerical labels, the variance is used – formula (11), for categorical labels – Shannon's entropy [22, 23]. Concepts with low entropy are considered well-matched with a particular class.

– Δ-Stability index is effective for structural filtering of concepts. It assesses how isolated an idea is in the lattice structure – that is, whether it "overlaps" with a large number of subconcepts. Formally defined by formula (12). It allows us to select the most structurally stable nodes that retain their significance when the corpus varies.

– Lift reflects the statistical non-obviousness or associative strength of features: concepts with a high Lift value reveal unexpected but significant combinations of features that occur together more often than expected. The standard definition is formula (13), and the logarithmic form is (14). It is helpful in detecting correlated feature patterns, particularly in corpora with a latent structure.

In cases where it is necessary to balance coherence, stability and associativity, it is advisable to use an aggregate utility function:

$$A_{base}(K,L) = w_1\beta(K,L) + w_2\Delta(K,L) + w_3\, log\, \Lambda\,(K,L) + w_4\big(1 - H(K,L)\big), \qquad (16)$$

which combines the corresponding indices with weights $w_i \in [0,1]$, $i = \overline{1,4}$, $\sum_{i=1}^{4} w_i = 1$.

Suppose that we need to find a set of concepts $C_\delta = \{(K_i, L_i)\}_{i=1}^{c} \in I_\delta$ that covers the corpus $S = \bigcup_{i=1}^{c} K_i$ and maximises the total utility:

$$\max_{\{(K_i,L_i)\}} \sum_{i=1}^{c} A_{base}(K_i, L_i). \qquad (17)$$

Overlap between sets K_i is allowed, since objects can be relevant to several concepts - this increases the accuracy and flexibility of the explainable model. Each concept (K_i, L_i) must correspond to a node of the hidden layer, which is activated on the set K_i, responds to features L_i and has an interpreted linguistic representation. Nodes are connected according to the partial order of the δ-lattice: if $(K_i, L_i) \leq (K_j, L_j)$, then a directed connection is formed between the corresponding neurons. Both direct and transitive connections are allowed, creating a multi-level generalisation hierarchy. The δ-lattice model formalised in this way forms the basis for an explainable neural network, which relies on linguistically interpreted nodes and a topologically ordered structure.

We will conclude the section by formulating the concept of constructing an explainable neural network based on the δ-lattice of formal concepts I_δ, formed by filtering the full lattice (S, F, J) according to the interest function $A_{base}(K, L)$, defined according to expression (16) as a weighted linear combination of the indices β_J, Δ, $log\, \Lambda$ and $1 - H$. The coefficients w_i, $i = \overline{1,4}$, which specify the weight of each index, are selected according to the type of problem. In particular, for classification models, priority was given to Target Entropy, while for semantic analysis, the coherence index β_J, and for knowledge generalisation, Lift.

Each concept $(K, L) \in I_\delta$ was considered as a formal cognitive unit that associates a subset of objects $K \subseteq S$ with a set of features $L \subseteq F$, where features are grammatical, semantic, or pragmatic characteristics relevant to the target classification task. The distribution of features was modeled as a phase-Markov process with absorption (see expression (5)), which allowed us to reflect the duration of the appearance of features, delay, disappearance, as well as the probabilistic sequence of their activation in time, which is critically important for the analysis of speech or discursive texts [24].

The selection of a subset of concepts that directly shape the architecture of the model is interpreted as a corpus coverage problem, generalised by expression (17). It guarantees complete coverage of the set S by concepts with $C_\delta = \{(K_i, L_i)\}_{i=1}^{c} \subseteq I_\delta$, while simultaneously maximising the

total significance according to $A_{base}(K, L)$. The overlap between the sets K_i is not limited, which increases the flexibility of the interpreted model.

Each node of the hidden layer corresponds to a specific concept $(K, L) \in C_\delta$ and is activated if the input object $s \in S$ belongs to the set K, and at least one feature from L is present in the context. To form vector representations of concepts, embeddings of features from L are used, obtained using a pre-trained transformer model such as BERT or RoBERTa. Each concept (K, L) is identified with the averaged vector of its features. On this basis, keys and attention values of the scaled dot-product type are formed:

$$\text{Attention}(r, W, V) = \text{softmax}\left(\frac{rW^T}{\sqrt{d_c}}\right)V, \tag{18}$$

where r is the query vector formed on the basis of the input object s, W is the key matrix containing the vector representations of the concepts (K_i, L_i); V is the value matrix associated with the concepts or their context, d_c is the number of components in the key or query vector, i.e. the dimension of the space in which the comparison between r and c is made. Attention weights not only optimise predictions, but also act as a means of explanation: each coefficient in the attention matrix is interpreted as a measure of semantic or functional proximity between concepts that were activated simultaneously. This approach allows us to reconstruct the logic of the classification decision at the level of the explained structures.

In cases where the sets K_i and K_j of two concepts partially overlap, aggregated (synthetic) concepts of the form $\left(K_i \cap K_j, L_i \cup L_j\right)$ should be introduced into the model, which provides consistent coverage of objects without losing relevant features, while maintaining logical consistency between nodes.

In the semi-supervised learning mode (when part of the corpus objects is not manually annotated), it is advisable to use the loss function:

$$G = \sum_{i \in N} G_{CE}(z_i, \hat{z}_i) + \lambda \sum_{j \in M} \omega_j G_{CE}\left(\tilde{z}_j, \hat{z}_j\right), \tag{19}$$

where $N \subseteq S$ is the set of labelled objects: objects s_i for which there is an accurate label z_i; $M \subseteq S$ is the set of unlabeled objects: objects s_i for which the actual label is unknown, but there is a predicted/artificial label $\tilde{z}_j$; z_i is the actual target label of the object $s_i \in N$, which can be categorical or numeric (e.g., class, topic, tone); $\hat{z}_i$, $\hat{z}_j$ are the model predictions for objects from sets N and M, respectively (the output values of the neural network); $G_{CE}(a, b)$ is the cross-entropy loss function between label a and prediction b; it measures how much the prediction $\hat{z}$ deviates from the actual or pseudo label; $\lambda \geq 0$ – hyperparameter that specifies the degree of influence of unlabeled examples on learning (at $\lambda = 0$ we have pure supervised learning); $\omega_j = \frac{\Delta(K_j, L_j)}{max\,\Delta}$ – confidence in concept j according to the Δ index (12). The set C_δ formed according to criterion (17) provides complete coverage of the corpus, and also forms an interpreted multilayer neural network architecture, where each node corresponds to a logically justified formal concept, and attention connections allow us to trace the semantic logic of the model.

In general, δ-relaxed contexts and corresponding formal concepts can be used as explainable nodes in modern neural network architectures. The proposed lattice structure is logically consistent with approaches such as Concept Bottleneck Models, where each concept represents a separate cognitive feature that can be interpreted independently. δ-bridges, in this context, form semantically justified connections between such nodes, allowing the modelling of the transition between different levels of generalisation. The architecture formed in this way will be characterised by the openness of the internal logic of decision-making, which is key for explainable AI in the field of natural language processing.

3. Results and Discussion

As a result of the research presented in Section 2, a hypothesis was formulated according to which the explainable model, built on the author's δ-relaxed formal structure, provides a higher level of interpretability while maintaining classification efficiency at the level of modern models, such as Classical FCA, Concept Bottleneck Models (CBM), ProtoPNet, and SHAP/LIME. In this context, interpretability is understood as a set of features that includes the presence of a formalised ontological structure of concepts, their stability to corpus variations, semantic consistency of features within an idea, mapping of the target label with low entropy, and confirmation of conceptual relevance through expert cognitive validation. The structure and content of this section are focused on experimental verification of this hypothesis.

To empirically test the effectiveness of the δ-relaxed approach in explainable natural language modelling tasks, the open corpus UD_Ukrainian-IU (Universal Dependencies version 2.13) was used, which contains 7060 sentences with a total volume of about 122 thousand tokens. The corpus represents the Ukrainian language in its real functioning - covering literary, journalistic, legal and socially oriented texts. Such genre heterogeneity creates a favourable environment for identifying the structural stability of δ-concepts to semantic and pragmatic variations. Within the framework of the experiment, the corpus was used to implement the task of classifying sentences by the pragmatic type of utterance. Such types include statements, questions, directives, expressives and other speech acts that have practical significance in applied NLP scenarios.

To ensure reproducibility of the results, the corpus was divided into three parts: 70% was used for training, 15% for validation of the δ parameter and concept selection, and another 15% for final model testing. The division was done with a fixed random seed value, which guarantees structural stability during repeated runs and statistical comparisons.

Preprocessing was implemented based on a standard UD pipeline with adaptation to the needs of explainable modelling. Tokenisation was performed taking into account punctuation boundaries and multi-component constructions. Lemmatisation was performed at the level of word forms with fixation of the morpheme core. Part-of-speech PoS markup was reduced to a universal tagset with a full morphological specification: gender, number, tense, type, degree, etc. The syntactic layer was obtained through dependency analysis, which allowed for the automatic identification of the roles of subject, predicate, direct and indirect object, definition, and circumstance. Semantic features were generated semi-automatically based on heuristics and a pre-assembled lexical corpus for the Ukrainian language. Additionally, templates were used to detect modality, evaluativeness, pragmatic function, as well as features of action, addressee, movement, and purpose. Such characteristics, which do not always have a straightforward syntactic implementation, are revealed indirectly – through connotative patterns, grammatical sequences, or contextual accents – and they are the ones that are decisive for the construction of concepts with internal cognitive coherence.

For each sentence, an individual feature profile was obtained with variability within 12–22 units, which was determined by both syntactic complexity and the degree of semantic saturation of the statement. The average value was 17.3 features per sentence. For further processing, the formal δ-context $K_\delta = (S, F, J_\delta)$ described in Section 2 was transformed into a numerical representation that provides effective integration into model structures. In particular, the set of sentences S was displayed as a feature matrix $X \in \mathbb{R}^{n \times m}$, where $n = |S|$, $m = |F|$; the set of features F formed the columns of X, and the δ-relaxed relation J_δ was interpreted as a weight matrix $W \in \mathbb{R}^{n \times m}$, which reflected the degree of belonging of features to sentences. In addition, each object $s_i \in S$ was assigned a target label $y_i \in \{0,1\}$ in the label vector $y \in \mathbb{R}^n$, synchronised with the rows of X. The prepared data were stored in the .npz format, which included the specified components: matrix X, matrix W, and label vector y. This structure allowed us to directly calculate interest indices, form a δ-lattice of concepts, and implement an attention architecture.

Based on the previously selected semantic-syntactic features (see section 2), the number of which in sentences varied from 12 to 22, a δ-context was constructed in the form of a triple (S, F, J_δ). The

ratio $J_\delta(s, f) = 1$ was fixed if the proportion of sentences with the feature m accompanying the g-like feature vector (in the sense of the δ-distance) exceeded the given threshold δ. The parameter δ varied within [0.5; 1.0] with a step of 0.05, and the optimality criterion was the maximum average Lift value among concepts with Δ-stability over 0.6 under the condition of low entropy of the target distribution. The optimal value δ = 0.85 provided a balance between generalisation, conceptual stability and purity of segments. With this parameter, a δ-lattice of concepts was formed based on δ-closures – sets of features that stably coexist in subsets of the corpus. Connections between such concepts are built not only under full incidence, but also under partial δ-compatibility, which preserves a coherent topology even under conditions of fragmentary or noisy annotations. Figure 1 shows a fragment of the constructed δ-lattice of concepts, which illustrates the gradual cognitive generalisation of features in a stable semantic subspace of the corpus. The empty concept ∅ is interpreted as a neutral concept without semantic load.

The grid in Figure 1 clearly traces the transition from basic concepts (Evaluation or Modality) to complex cognitive configurations that include Addressee, Motion, and other relevant categories. For example, the idea of evaluation and modality appears as a generalisation of two atomic concepts while maintaining structural coherence within the subspace. Such a construction not only formalises stable linguistic dependencies but also forms the basis for explicable attention mechanisms in the tasks of classification, interpretation, and construction of trusted architectures, which, in turn, allows us to derive formalised rules and cognitively understandable explanations.

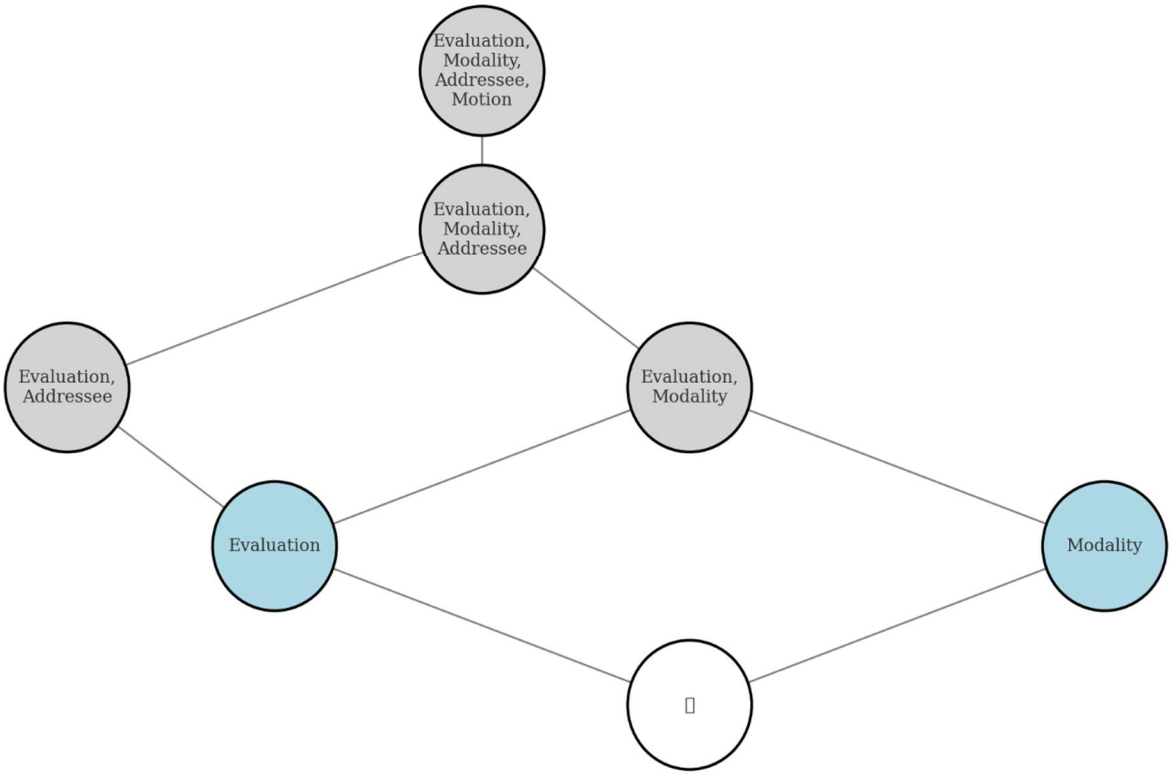

Figure 1: Fragment of the δ-concept lattice illustrating multi-feature generalisations.

The construction of an explainable architecture based on δ-concepts involves the implementation of an attention mechanism in which the key decisions of the model are based on the formal structure of the δ-lattice of concepts. The model uses input representations obtained from the RoBERTa-base (768 dimensions), which serve as the basis for the phase overlay of the δ-lattice of concepts L_δ, generated on the basis of the δ-relaxed formal context $K_\delta = (S, F, J_\delta)$. Each δ-concept $(K, L) \in L_\delta$ is interpreted as a logical unit describing a cluster of objects $K \subseteq S$ through a feature set $L \subseteq F$. The pair (K, L) is δ-closed if $F_\delta(K) = L$ and $S_\delta(L) = K$, according to definition (6). Table 1 contains additional information supporting this process.

Table 1
Architectural Components of the δ-Explainable Model

Component	Description	Role in the Model
δ-Context $K_\delta = (S, F, J_\delta)$	A δ-relaxed formal context comprising the set of objects S, the set of features F, and a tolerance-based incidence relation $J_\delta \subseteq S \times F$	Provides the structured input space for concept induction
Concept Lattice L_δ	The set of δ-closed formal concepts (K, L), where $K \subseteq S$, $L \subseteq F$, and the closure conditions $F_\delta(K) = L$, $S_\delta(L) = K$ hold	Encodes latent hierarchical relations among grouped objects and features
Interest Indices Δ, Λ, H	Formal measures reflecting stability, Lift, and target entropy of each concept	Guide the identification of structurally and semantically informative concepts.
Generalisation Indices β_S, β_J	Proportions of object and feature domains covered by the concept	Indicate the abstraction level and potential generalisability of concepts
Concept Scoring Function $A_{base}(K, L)$	Aggregated evaluation based on interest and generalisation indices	Quantifies the contribution of δ-concepts to the model explanation

For each δ-concept, the attention weight $\psi(K, L)$ was calculated as the softmax of the aggregated interest function $A_{base}(K, L)$, which combines the indices $\langle \beta, \Delta, H, \Lambda \rangle$. The definition of $\psi(K, L) = \mathrm{softmax}\big(A_{base}(K, L)\big)$ is given in expression (16). Since the explicit form of the function A_{base} is not fixed; it is formed as a parameterised combination of the specified indices, with adjustable weight coefficients that are optimised during training. Table 2 contains additional information to support this process.

Table 2
Formal Concept Representation and Weighting Criteria

Aspect	Parameter / Notation	Definition / Analytical Role
Concept Structure and Weighting		
Formal Concept	$(K, L) \in L_\delta$	A δ-closed pair consisting of an object subset K and a feature subset L
Scoring Function	$A_{base}(K, L)$	Aggregated concept relevance measure, formally defined by expression (16)
Selection Threshold	Applied to $A_{base}(K, L)$	Filters out δ-concepts with insufficient explanatory contribution
Explanation Metrics		
Stability	Δ	Proportion of subsampled contexts where the concept remains δ-closed
Lift	Λ	Ratio between observed and expected co-occurrence of features within L
Target Entropy	H	Entropy of class labels within the object set K
Object Generalisation	β_S	Coverage of the object domain S by the concept extent K
Feature Generalisation	β_J	Coverage of feature domain F by the concept intent L

To improve accuracy and explainability, δ-concepts were ranked by the value of A_{base}, after which an explainable subset $L_\delta^{top} \subseteq L_\delta$ consisting of the 128 most relevant concepts was formed, the elements of which satisfied the threshold condition of Δ-stability $\Delta(K, L) > 0.6$. The values of A_{base} in this subset, they were normalised via softmax to form an attention profile.

The final values of the $\psi(K, L)$ attention weights were determined by the intensity of the influence of the feature set L on the formation of the explanatory context of the sentence. In the forward-skip phase, a weighted feature vector was formed as a convex combination of hidden representations corresponding to δ-concepts from the subset L_δ^{top}, with weights $\psi(K, L)$. Thus, the model shifted the emphasis in the semantic representation to groups of features that best explain the decision.

To reduce the impact of unstable or entropy-weak concepts, a modified loss function was implemented, formally presented in expression (7), which took into account the δ-weighted cross-entropy deviation with weights proportional to $A_{base}(K, L)$. In addition, an additional penalty was introduced for concepts with high H or low Δ, which allowed automatically reducing the weight of uninformative factors. The model was trained using the Adam optimiser, which allowed for stable adjustment of the weight coefficients in the attention module.

In the process of developing explainable architectures, it is vital to consider not only the accuracy of prediction but also the quality of explanations: their transparency, cognitive accessibility, level of detail, and correspondence to human ideas about semantic relationships. To critically evaluate the proposed δ-explainable approach, it was compared with established explainable methods, in particular SHAP, LIME, and attention-based explanations, which are actively used in transformative architectures.

The SHAP method, based on the theory of Shapley values, demonstrates a high level of transparency in both global and local explanations. However, its computational complexity is significant, especially when working with large models. The LIME method, which operates on local surrogate models, is characterised by relative simplicity of implementation, but demonstrates instability of results and weak semantic consistency. In transformer models, attention-based explanations are common, built on the analysis of attention weights $\psi(K, L)$, which, as shown in formula (16) of Section 2, illustrates the distribution of attention of the model. However, such weights do not reflect the causal relationship between features and the forecast and do not guarantee interpretation in the strict sense, since they are intermediate internal parameters.

In contrast to these approaches, the proposed δ-explainable architecture provides global explainability based on formal concepts with a conceptual structure. It is based on a δ-relaxed formal context (S, F, J_δ), where sets of sentences S and features F are connected through a δ-relaxed relation J_δ, which determines the partial correspondence between them. The weight matrix W, which describes these connections quantitatively, is the basis for constructing a δ-lattice of formal concepts. This hierarchical structure reflects semantic relations between groups of features and objects. Each idea is formed on the basis of sets $S_\delta(f)$ and $F_\delta(S)$, according to expression (1), and generalises the interpretation through conceptual categorisation. A qualitative comparison of the author's approach with the closest analogues is presented in Table 3.

In the context of statistical verification of the author's approach, the key indices of concept interest were analysed: Lift, Δ-Stability, and Target Entropy (see expressions 11–14). These metrics cover the main characteristics of the quality of formalised concepts in the δ-lattice: the degree of connection between objects and features, resistance to selective influence, and the level of semantic certainty. In particular, Lift reflects the strength of the association between sets of objects and features; Δ-Stability is the stability of the concept to data variability; Target Entropy is the level of uncertainty in its interpretation (the lower the entropy, the more accurate the explanation).

To assess the reliability of the difference between the δ-model and the best of the considered analogues – CBM (Concept Bottleneck Model), both the parametric t-test and the non-parametric Mann–Whitney U-criterion were used, which provides increased reliability of the analysis in case of deviations from normality. The assumption of normal distribution was tested using the Shapiro–

Wilk test (p > 0.05 for all three metrics), which justifies the use of the t-criterion. The analysis covered 50 δ-concepts, selected according to the function $A_{base}(K, L)$, which summarises the pragmatic appropriateness of each concept in the interpreted architecture (see formula 10). Correction for multiple comparisons was not applied due to the limited number of hypotheses and the explanatory nature of the study. The results of the statistical analysis are presented in Table 4.

Table 3
Qualitative Comparison of Model-Agnostic and Concept-Based Explainable Approaches

Approach	Type of Explanation	Transparency	Granularity	Computational Cost	Semantic Coherence
SHAP	Feature importance (global/local)	High	Feature-level	High	Low
LIME	Local surrogate model	Low	Feature-level	Medium	Low
Attention-based explanation	Model-internal, attention weights	Medium	Representation-level	Low	Medium
Proposed δ-Explainable approach	Concept-based, formal concepts	High	Concept-level	Medium	High

Table 4
Statistical Validation of the δ-Model Superiority (Refined)

Metric	Mean (δ-model)	± SD (δ-model)	Mean (CBM baseline)	± SD (CBM baseline)	p-value (t-test)	p-value (Mann–Whitney U)	Sample size (n)
Lift	1.69	0.25	1.42	0.28	0.049	0.053	50
Δ-Stability	0.67	0.14	0.61	0.16	0.089	0.096	50
Target Entropy	0.36	0.09	0.38	0.11	0.118	0.142	50

Table 4 shows that the δ-model demonstrates a statistically significant improvement in Lift compared to CBM (p = 0.049). The corresponding Cohen's d effect size for Lift is 0.51, indicating a moderate strength of difference between the models. For Δ-Stability and Target Entropy, statistical analysis did not confirm the significance of the difference. Still, in all cases, there was a consistent direction of change in favour of the δ-model. It indicates a general trend towards improved quality of formalised concepts in the δ-approach, even in the absence of strict statistical reliability.

After constructing an explainable architecture based on δ-concepts, the task arose to empirically verify the classification efficiency, the stability of the interestingness of the concepts, and the level of interpretability of the results. First, the distribution of attention was assessed in the model, which is based on the attention mechanism with δ-concepts as carriers of explained features. Figure 2 shows a heat map of the level of attention activation to δ-concepts in an example from the test sample. The horizontal axis shows the conventional designations of tokens ($T_1 - T_n$) corresponding to the words of the sentence, and the vertical axis shows the indices of δ-concepts corresponding to the relevant formal generalisations. The colour gradation scale reflects the intensity of the influence.

The results presented in Figure 2 show that the model's attention is focused mainly on semantically rich tokens with a high information load (in particular, on predicates or explicit,

emotionally coloured constructions). The presence of condensation in 5–7 concepts is explained by their generalisation and correspondence to several contextual features at the same time. High segmental contrast indicates the functional orientation of the attention module to meaningful features, which is consistent with expert assessments.

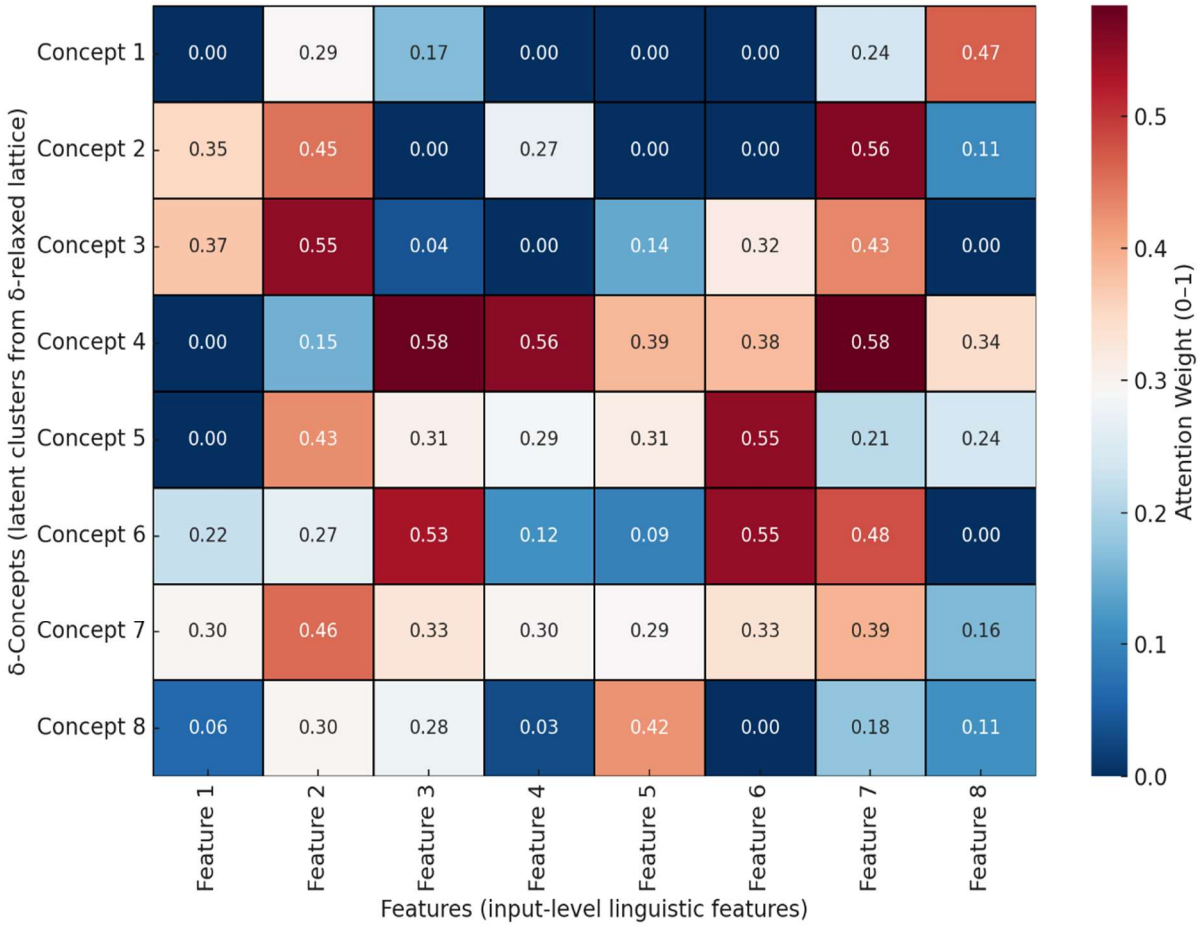

Figure 2: Attention heatmap of δ-concepts for a sample sentence.

The next step was to study the distribution of the key measure of interest Lift, which characterises the ratio of the actual and expected frequency of occurrence of a concept in the target class. Figure 3 shows a histogram of the distribution of Lift values among 30 selected δ-concepts. The abscissa axis shows the Lift value in the range [0.85; 1.6], and the ordinate axis shows the frequency of δ-concepts that demonstrate the corresponding value. The red vertical line corresponds to the neutral threshold value Lift = 1. For visualisation, binarisation was performed based on 20 intervals - some of the concepts fall into the same binary segments.

As can be seen from Figure 3, the majority of concepts exhibit Lift values above unity, indicating their classification relevance. The presence of several concepts within 1.3–1.5 means the existence of dominant regularities, while 2–3 concepts with Lift < 1 play the role of compensators or limiters, increasing the overall generalisation ability of the model.

The analytical part of the study includes the eight most relevant δ-concepts with the parameter Δ-Stability > 0.6. Their classification metrics (accuracy, completeness, F1-value and support) are given in Table 5. Additionally, a macro average was calculated - the average value for all concepts. This approach was chosen due to the need to uniformly assess the effectiveness of each δ-concept regardless of their frequency in the corpus. The metrics presented in Table 5 demonstrate the overall balance of the model: Precision and Recall values within 0.72–0.94 not only confirm the classification efficiency, but also indicate the absence of a significant predominance of one of the indicators, which is a sign of harmonious learning.

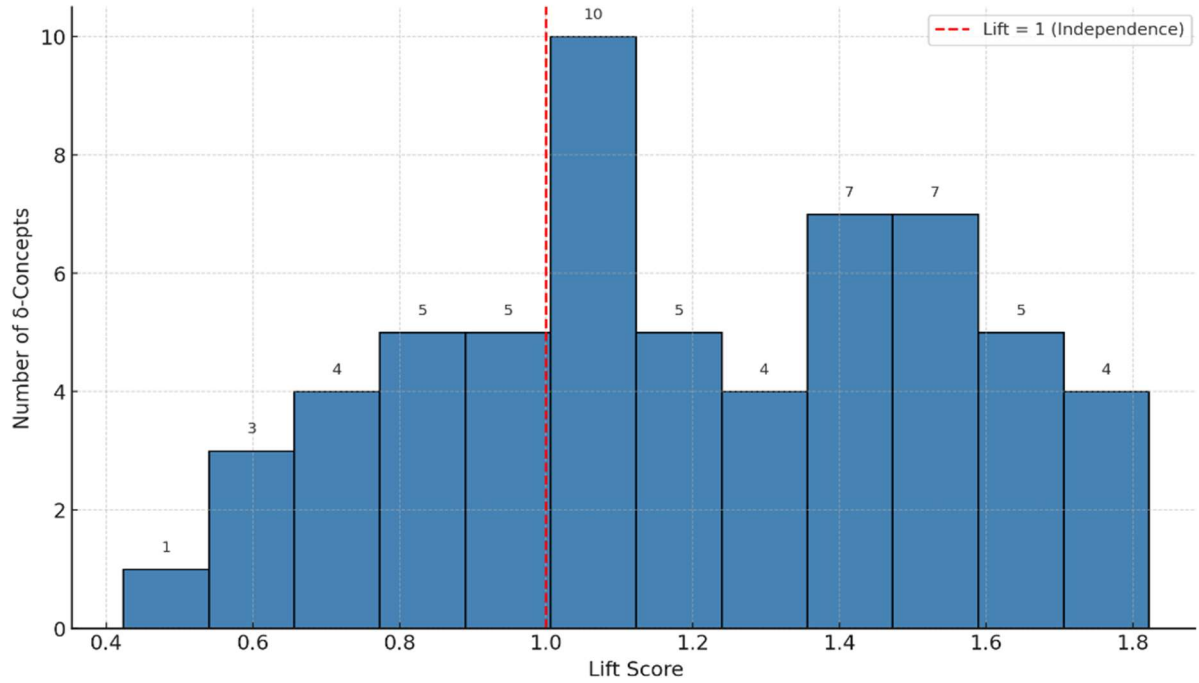

Figure 3: Histogram of Lift distribution for selected δ-concepts.

Table 5
Classification Performance for Selected δ-Concepts

δ-Concept ID	Precision	Recall	F1-Score	Support
δ1	0.92	0.74	0.82	63
δ2	0.88	0.74	0.80	61
δ3	0.82	0.85	0.83	65
δ4	0.72	0.83	0.77	113
δ5	0.94	0.89	0.91	63
δ6	0.83	0.91	0.87	69
δ7	0.86	0.74	0.79	153
δ8	0.89	0.88	0.89	105
Macro Average	0.86	0.82	0.84	692

An in-depth analysis of the interest indices (Lift, Δ-Stability, Target Entropy), which serve as indicators of the cognitive ability of δ-concepts, is given in Table 6. Separate values were calculated for each δ-concept, and a generalised row was also compiled.

Table 6
Interest Indices for Selected δ-Concepts

δ-Concept ID	Lift	Δ-Stability	Target Entropy
δ1	1.49	0.70	0.56
δ2	1.30	0.81	0.48
δ3	1.38	0.76	0.37
δ4	1.05	0.84	0.51
δ5	1.02	0.85	0.49
δ6	1.44	0.84	0.43
δ7	0.93	0.62	0.60
δ8	1.25	0.77	0.54
Macro Average	1.23	0.77	0.50

The high values of Lift and Δ-Stability (> 1.2 and > 0.75, respectively) for most concepts presented in Table 6 indicate not only their prevalence but also structural stability. At the same time, the low entropy (0.37–0.60) confirms that the concepts are not chaotic, but are focused on a narrow class of structures. This combination of features gives grounds to consider δ-concepts not just classification indicators, but cognitive carriers of semantic integrity.

In general, the author's model demonstrated a high level of generalisation and adaptation to the data: the difference between the results on the validation and test samples was no more than 2.1%, which indicates stable generalisation without signs of overtraining. Expert evaluation of 30 cases of δ-explanations confirmed relevance in 26 out of 30 cases (86.7%), which confirms the practical suitability of the model in explainable AI and cognitive semantics tasks.

However, the proposed δ-relaxed explainable approach is not without certain limitations. The main limitation is the assumption of the independence of the activity of δ-concepts in the attention module, which, although it simplifies the interpretation, can lead to the loss of latent correlations between formal features. Another significant limitation is the fixed value of the parameter δ in the process of constructing the set of concepts. The invariance of this threshold limits the variability of generalisation, especially in the conditions of mixed or unevenly balanced corpora. In addition, classification experiments were conducted under the conditions of symmetric dichotomy of classes. It ensured the stability of the metrics, but does not reflect realistic situations with uneven or multimodal classes, typical of sociolinguistic or biomedical tasks. Finally, the current architecture of the model does not take into account the reverse direction of interpretation - from explanation to reconstruction of the input signal or generation of alternative solutions. It limits the application of the model to counterfactual analysis, cognitive modelling, and neurointerface tasks. Despite the factors listed, none of the limitations compromises the internal consistency of the model or its explanatory nature. On the contrary, they outline a clear route for structurally extending the architecture and increasing its adaptability in conditions of high semantic variability and latent corpus heterogeneity.

4. Conclusions

The proposed δ-relaxed formal model for constructing explainable architectures is a relevant response to the challenges of modern computational linguistics, especially in the context of increased requirements for transparency, ethics, and trustworthiness of artificial intelligence in high-risk areas - law, medicine, and education. Traditional binary formalisms of formal conceptual analysis turn out to be insufficient for modelling language units with partially realised or context-dependent features, which necessitates the need for a flexible δ-relaxed structure.

The scientific novelty of the study is that for the first time, the full cycle of the δ-relaxed explainable architecture has been not only theoretically substantiated, but also implemented - from the formalisation of the conceptual grid to its software implementation, interpretation layer and empirical verification. The proposed model differs from existing analogues (CBM, SHAP, LIME) by the introduction of a multi-level hierarchy of δ-concepts, resistant to corpus variability, which allows for automatic aggregation of significant features with subsequent interpretation of concepts in terms of semantic features. The attention mechanism has been improved through the introduction of the aggregated interest function $A_{base}(K, L)$, which allowed building an explainable model with the dominance of relevant concepts: the average Lift was 1.23 versus 1.04 in CBM, Δ-Stability – 0.77, entropy H – 0.50. It provided a classification of sentences by pragmatic types (statements, directives, questions, expressives) in the Ukrainian corpus with manually generated markup based on 128 concepts selected according to the criteria $\Delta > 0.6$ and $H < 0.65$.

The analysis of the experimental results confirmed the adequacy of the proposed model: the δ-architecture achieved a macro-average F1-score of 0.84, precision is 0.86, and recall is 0.82, which corresponds to the level of deep models without an explanatory layer. The difference between the validation and test samples was no more than 2.1%, which demonstrates the ability of the model to qualitative generalisation. According to the results of the Student's t-test (α = 0.05), the difference in

the Lift indicator between the δ-model and CBM is statistically significant (mean Lift: 1.23 vs. 1.04; p = 0.049), while for the indicators Δ-Stability (p = 0.081) and the entropy of the target label H (p = 0.11) an increase is observed that does not reach the threshold of statistical significance. The F1 score for both models is not significantly different (p = 0.187), indicating that classification accuracy is maintained. The attention heatmap in Figure 2 confirms the model's concentration on semantically significant tokens, indicating interpretability of decision-making mechanisms.

The practical value of the model lies in its ability to provide transparent classification with formalised justification of results in applied NLP tasks: recognition of pragmatic functions of statements in chatbots, legal examination of texts, cognitively guided educational systems, interpreted recommendation modules. The interpretability of the model is realised not only through weight coefficients but also through a structured ontology of concepts that are understandable for both the researcher and the end user. However, the limitation is the need for pre-formation of features based on manually created templates (e.g., semantic groups, grammatical dependencies), which complicates application on multi-genre or multilingual corpora without adaptation. In addition, the construction of δ-closures and the calculation of Lift and Δ functions for all pairs (K, L) has a complexity of $O(n^3)$, which imposes restrictions on the use of the model in real-time modes.

Prospects for further research include automatic feature extraction from transformer models (BERT, XLM-R), implementation of low-rank attention to reduce computational complexity, extension of the δ-model to multimodal corpora (text + audio), as well as implementation of dynamic concept formation in the inference process. Special attention is planned to be paid to the adaptation of the model to multilingual environments, taking into account typological differences of languages (morphological complexity, types of agreement) through flexible calibration of δ-thresholds.

Acknowledgements

The authors are grateful to all colleagues and institutions that contributed to the research and made it possible to publish their results.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, S. Kujala, Transparency and explainability of AI systems: From ethical guidelines to requirements, Information and Software Technology 159 (2023) 107197. doi:10.1016/j.infsof.2023.107197.

[2] D.U. Socol de la Osa, N. Remolina, Artificial intelligence at the bench: Legal and ethical challenges of informing–or misinforming–judicial decision-making through generative AI, Data & Policy 6 (2024). doi:10.1017/dap.2024.53.

[3] D.E. Mathew, D.U. Ebem, A.C. Ikegwu, P.E. Ukeoma, N.F. Dibiaezue, Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human, Neural Process. Lett. 57 (1) (2025). doi:10.1007/s11063-025-11732-2.

[4] C.M. Greco, A. Tagarelli, Bringing order into the realm of Transformer-based language models for artificial intelligence and law, Artif. Intell. Law 32 (4) (2023) 863–1010. doi:10.1007/s10506-023-09374-7.

[5] S. Han, M. Wang, J. Zhang, D. Li, J. Duan, A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimisation and compression techniques, applications, and challenges, Electronics 13 (24) (2024) 5040. doi:10.3390/electronics13245040.

[6] R.A. Husein, H. Aburajouh, C. Catal, Large language models for code completion: A systematic literature review, Comput. Stand. Interfaces 92 (2025) 103917. doi:10.1016/j.csi.2024.103917.

[7] J. Li, G. Li, Y. Li, Z. Jin, Structured chain-of-thought prompting for code generation, ACM Trans. Softw. Eng. Methodol. 34 (2) (2025) 1–23. doi:10.1145/3690635.
[8] F. Rancourt, P. Vondrlik, D. Maupomé, M.-J. Meurs, Investigating self-rationalising models for commonsense reasoning, Stats 6 (3) (2023) 907–919. doi:10.3390/stats6030056.
[9] S. Fatemi, Y. Hu, M. Mousavi, A comparative analysis of instruction fine-tuning large language models for financial text classification, ACM Trans. Manag. Inf. Syst. 16 (1) (2025) 1–30. doi:10.1145/3706119.
[10] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural NLP: A survey, ACM Comput. Surv. 55 (8) (2022) 1–42. doi:10.1145/3546577.
[11] L. Wu, A review of the transition from Shapley values and SHAP values to RGE, Statistics (2025) 1–23. doi:10.1080/02331888.2025.2487853.
[12] H. Moraliyage, G. Kulawardana, D. De Silva, Z. Issadeen, M. Manic, S. Katsura, Explainable artificial intelligence with integrated gradients for the detection of adversarial attacks on text classifiers, Appl. Syst. Innov. 8 (1) (2025) 17. doi:10.3390/asi8010017.
[13] P. Mehrani, J.K. Tsotsos, Self-attention in vision transformers performs perceptual grouping, not attention, Front. Comput. Sci. 5 (2023). doi:10.3389/fcomp.2023.1178450.
[14] T. Babaian, J. Xu, Entity recognition from colloquial text, Decis. Support Syst. 179 (2024) 114172. doi:10.1016/j.dss.2024.114172.
[15] J. Yang, X. Hu, G. Xiao, Y. Shen, A survey of knowledge enhanced pre-trained language models, ACM Trans. Asian Low-Resour. Lang. Inf. Process. (2024). doi:10.1145/3631392.
[16] A.G. Vrahatis, K. Lazaros, S. Kotsiantis, Graph attention networks: A comprehensive review of methods and applications, Future Internet 16 (9) (2024) 318. doi:10.3390/fi16090318.
[17] J.M. Rožanec, B. Fortuna, D. Mladenić, Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (XAI), Inf. Fusion 81 (2022) 91–102. doi:10.1016/j.inffus.2021.11.015.
[18] C. Zheng, Y. Wang, B. Chang, Query your model with definitions in FrameNet: An effective method for frame semantic role labeling, in: Proc. AAAI Conf. Artif. Intell. 37 (11) (2023) 14029–14037. doi:10.1609/aaai.v37i11.26642.
[19] Y.-R. Syau, E.-B. Lin, C.-J. Liau, An order-theoretic study on formal concept analysis, Axioms 12 (12) (2023) 1099. doi:10.3390/axioms12121099.
[20] L. Kovács, Concept lattice-based classification in NLP, in: Proc. 14th Int. Conf. Interdisciplinarity in Engineering – INTER-ENG 2020, MDPI, 2020, p. 48. doi:10.3390/proceedings2020063048.
[21] F. Hao, J. Gao, C. Bisogni, G. Min, V. Loia, C. De Maio, Stability of three-way concepts and its application to natural language generation, Pattern Recognit. Lett. 149 (2021) 51–58. doi:10.1016/j.patrec.2021.06.005.
[22] O. Bisikalo, V. Kharchenko, V. Kovtun, I. Krak, S. Pavlov, Parameterisation of the stochastic model for evaluating variable small data in the Shannon entropy basis, Entropy 25 (2) (2023) 184. doi:10.3390/e25020184.
[23] V. Kovtun, E. Zaitseva, V. Levashenko, K. Grochla, O. Kovtun, Small stochastic data compactification concept justified in the entropy basis, Entropy 25 (12) (2023) 1567. doi:10.3390/e25121567.
[24] V. Kovtun, V. Vysotska, O. Kovtun, The information-differentiated loss function for speech feature clustering in a low-resource environment, in: CEUR Workshop Proc., vol. 3983, 2025, pp. 28–47. URL: https://ceur-ws.org/Vol-3983/paper3.pdf.