

Context-Aware Semantic Fusion for Mathematical Formula Equivalence Detection

Andrii Dyriv*

Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine

E-mail: andrii.i.dyriv@lpnu.ua

ORCID iD: <https://orcid.org/0009-0009-0615-1161>

*Corresponding Author

Olga Lozynska

Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine

E-mail: olha.v.lozynska@lpnu.ua

ORCID iD: <https://orcid.org/0000-0002-5079-0544>

Victoria Vysotska

Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, 79013, Ukraine

Combating Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, 61080, Ukraine

E-mail: Victoria.A.Vysotska@lpnu.ua

ORCID iD: <https://orcid.org/0000-0001-6417-3689>

Dmytro Uhryn

Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine

E-mail: d.ugryn@chnu.edu.ua

ORCID iD: <https://orcid.org/0000-0003-4858-4511>

Yuriy Ushenko

Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine

E-mail: y.ushenko@chnu.edu.ua

ORCID iD: <https://orcid.org/0000-0003-1767-1882>

Received: August 28, 2025; Revised: February 25, 2026; Accepted: March 28, 2026; Published: August 8, 2026

Abstract: The paper investigates the problem of automatically detecting equivalent mathematical formulas in scientific texts. The authors propose a novel hybrid approach that combines structural analysis of formulas (normalisation, Abstract Syntax Tree (AST) construction, and vectorisation) with deep semantic analysis of the surrounding publication text using transformer models such as SciBERT and Sentence Transformers. During the study, a software implementation was developed that uses cosine similarity to assess context proximity and a Siamese neural network for equivalence classification. Evaluated on a custom dataset of 12,500 formula-context pairs from academic papers, the proposed hybrid model achieved an F1-score of 0.88, significantly outperforming baseline models that rely solely on structural (F1: 0.71) or textual (F1: 0.59) features. The experimental results were further visualised using heat maps, dendrograms, and UMAP projections, confirming the model's ability to identify equivalent expressions even when their syntactic notation differs significantly. The proposed framework is promising for use in anti-plagiarism systems, intelligent search services, and digital scientific libraries.

Index Terms: Mathematical formulas, semantic analysis, equivalence of formulas, publication context, vector representation (Embeddings), transformer models (SciBERT, MathBERT), machine learning, clustering, natural language processing (NLP)

1. Introduction

Automatic detection of equivalent mathematical formulas in scientific texts is one of the key and urgent tasks of modern information science, due to the rapid growth of electronic publications, the need to organise knowledge, the fight

against plagiarism, the search for scientific ideas, and even support for the educational process in higher educational institutions. The universalisation of access to large digital libraries, the increase in the number of publications with complex mathematical expressions and formulas, highlights the urgent need to create intelligent search systems that are able not only to find identical lexical or structural records, but also to recognise equivalent formulas regardless of how they are presented in the text. This task is critically important for analytical systems in science, education, and patents, as it increases transparency in scientific circulation and supports open access to knowledge.

An analysis of existing approaches in this area demonstrates that most classical systems focus either on the syntactic analysis of formulas (i.e., on the transformation of mathematical expressions into trees or graphs with subsequent lexical or structural comparisons) or on the comparison of formulas after bringing them to a certain canonical form by means of symbolic calculations. Such technologies include structuring formulas in MathML, using AST (Abstract Syntax Tree), applying various normalisation algorithms (e.g., symbolic simplification in SymPy), and approaches based on substring matches or heuristic variable substitution. However, these methods have limitations: they often "do not see" the semantic proximity between, for example, different but equivalent entries (for example, $a^2 + 2ab + b^2$ and $(a + b)^2$), do not take into account the context in which this formula was given, and do not "understand" the author's load of certain symbols or structure within a particular work. In many cases, changing the order of terms, notation, indices, or even the use of synonyms (e.g., x instead of a) causes formulas to be omitted as non-equivalent, even though they essentially mean the same thing. There is also another problem: a significant increase in accuracy often requires a deep semantic interpretation, which can be grounded in the meanings the formula acquires in a real-world context (e.g., physics or statistics) or in its role in the proof.

The emergence of new approaches has marked recent years, focusing on the use of natural language processing tools and modern machine learning architectures, in particular, embeddings, transformer models, and multilevel mechanisms for combining text and structure. At the same time, the problem of taking context into account remains insufficiently addressed: a formula can be equivalent only within the theoretical framework or application described in the publication (for example, the same algebraic formula in physics and chemistry can have different meanings).

That is why this work aims to develop and experimentally investigate a new approach to the automatic detection of equivalent mathematical formulas, combining semantic analysis of the formulas themselves with consideration of the contextual features of the publications in which they appear. To achieve this, the study sets itself the following main tasks: to analyse modern approaches to equivalence recognition, to develop a hybrid model that includes vectorisation of formulas along with embeddings of the surrounding text, to conduct an experimental comparison of different strategies for integrating context, to evaluate the quality and advantages of the new approach using classical metrics (accuracy, completeness, F1-measure) on open and specialised corpora. The expected scientific contribution is to prove the prospects of multicomponent analysis, which takes into account both the formal structure of formulas and their semantic space in the text, as well as to create a universal framework suitable for extension and use in problems of scientific search, automatic classification and assessment of plagiarism in mathematical texts.

2. Related Work

The current state of development of information technologies for processing scientific texts is characterised by a shift from purely syntactic methods to deep semantic analysis. The problem of identifying mathematical content covers several key areas: detecting expressions, analysing their internal structure, and checking for plagiarism. The main areas of research are the following tasks:

1. Detection and segmentation of mathematical expressions
2. Semantic equivalence and autoformalization
3. Specifics of natural language processing and intellectual analysis
4. Fighting plagiarism in the scientific community

The first stage of processing is the correct separation of formulas from the general text. The paper [1] provides a comprehensive analysis of methods for segmenting images of mathematical expressions, a critical step in the digitisation of printed publications. At the same time, [2] presented FormulaNet – a reference dataset that sets standards for training formula detection models.

The key task of our study is to recognise equivalent formulas. [3] propose an approach to autoformalization through a combination of symbolic equivalence and semantic consistency that correlates with your chosen strategy for using SymPy and BERT. The mathematical theory of semantic communication, described in [4], provides the fundamental foundation for the transmission and interpretation of the semantic load of complex systems.

Contextual similarity is also an important aspect [5-15]. The authors in [7] demonstrate the effectiveness of latent semantic analysis and cosine similarity for detecting borrowings. The authors in [10] focus on recognising formula concepts using machine learning, which allows the system to "understand" the purpose of the expression. The authors of [15] develop this idea by proposing LDA-based plagiarism detection systems that combine syntactic and semantic features.

Since the formula's context is often presented in natural language, linguistic analysis plays a crucial role. The authors

in [11-12] consider intelligent methods for processing Ukrainian content and mathematical models for decision-support systems for error correction, which are relevant to the localisation of such systems.

The study [13] on detecting paraphrases via binary classification of text pairs closely mirrors your approach of using Siamese architectures to compare object similarities.

The problem of "mathematical plagiarism" is more complex than the textual one. The authors in [8] propose a taxonomy of such plagiarism, emphasising the need to distinguish between identical entries and semantically equivalent ideas. Analytical studies in [10] confirm that traditional approaches are inferior to intellectual ones in identifying hidden borrowings.

The analysis of the literature shows that combining symbolic calculations to verify formal identity with transformer models for analysing the textual environment is the most promising direction for developing systems for scientific search and for evaluating the originality of works.

The relevance of the problem of automatically identifying equivalent mathematical formulas has led to the development of a wide range of methods for representing them in computer systems. There are three main paradigms in the literature: structural, symbolic and semantic. Structural approaches represent formulas as trees (parse trees, syntactic trees, or Directed Acyclic Graphs), where each node corresponds to an element or operation, and the tree explicitly preserves the hierarchy of operations.

For example, the expression $(a + b)^2$ after parsing is converted into a tree with an upper exponentiation operation and a subtree of addition $a + b$. In the structural approach, any comparison of formulas reduces to a comparison of tree isomorphisms. The standard for representing formulas in such systems is MathML or OpenMath. Table 1 compares classical approaches.

Table 1. Examples of Presentation of Formulas by Standards

Method	Submission format	Example	Feature
Structural and	MathML, AST	$(a + b)^2 \rightarrow \text{tree}$	Guarantees the preservation of the structure; sensitive to permutations and name changes
Symbolic	Symbolic expression	$(a + b)^2$	Allows canonical transformations; does not always distinguish between synonyms
Semantic and	Embedding, Earl	Formula vector	Takes into account meaning and context; needs ML/AI models

Symbolic approaches focus on analysing the original notation or transforming the formula into a certain canonical form. Modern symbolic libraries (e.g., SymPy, Mathematica) operate with the rules of algebraic simplification, which allows you to unify non-standard notations. Symbolic simplification makes it possible to understand that $a^2 + 2ab + b^2 = (a + b)^2$; however, it loses details of the structure of the view or local semantic nuances.

Semantic approaches go beyond structural isomorphism: the formula is transformed into a multidimensional vector space, where proximity is determined not only formally but also by the context of use. Semantic proximity of two formulas f_1, f_2 can be described in terms of the cosine distance between their vectors:

$$c(f_1, f_2) = (f_1 \cdot f_2) / (\|f_1\| \|f_2\|). \quad (1)$$

At the same time, in modern research, the question of the semantic similarity of the textual part, i.e., the place where the purpose, proof, or use of the formula is described, is of increasing interest. Methods such as Sentence Transformers, SciBERT, and SBERT allow you to build vector representations of entire paragraphs or sentences and thus determine the proximity of the contexts of formulas.

$$\text{sim}_{\text{conte}}(C_1, C_2) = \cos(c_1, c_2) = (c_1 \cdot c_2) / (\|c_1\| \|c_2\|). \quad (2)$$

To adapt such approaches to mathematical formulas, preliminary vectorisation or syntactic breakdown of expressions is required. In the literature, there are hybrid methods where embedding is formed as a concatenation of the MathML structure and the text environment:

$$F = [\text{Embed_MathML}(f); \text{Embed_BERT}(\text{Contextallows})]. \quad (3)$$

Comparison of different approaches in Table 2 to see their effectiveness in typical problems. The analysis of scientific sources shows that structural-symbolic models perform well at finding identical records, but their flexibility decreases significantly when records vary or when notation changes. Conversely, modern semantic embeddings combined with context analysis enable the identification of chaotically stated formulas that are semantically equivalent, thereby making the search more robust to formal differences. However, even the best of the studied approaches face difficulties in a multidisciplinary context – the same formula can have different interpretations across different sciences or even within the same work.

Table 2. Comparison of approaches

Model/Study		Task	Context integration feature	Precision	Recall / Completeness
MathBERT (Peng et al.)		Math search	Vectorisation of a formula and its text description	0.90	0.85
Tangent-CFT		Formula search	Combination of structural and textual features	0.82	0.77
Hybrid SymPy+BERT		Equivalence	Canonical representation of a formula with context embedding	0.87	0.72
Classic isomorph.	tree	Structural only	Only works with the structure of the formula without taking into account the context	0.60	0.61

While recent state-of-the-art models from 2024–2026, such as MathBERT and advanced contrastive learning frameworks, have significantly improved formula embeddings, they predominantly focus on internal token relationships within formulas. A critical limitation of these contemporary approaches is their shallow integration of the surrounding textual narrative. They often treat context as a secondary feature concatenated at the final layer, rather than employing a deep cross-modal attention mechanism to align mathematical variables with their explicit textual definitions dynamically. Overall, the literature review demonstrates a clear shift towards consolidating structural analysis with context-based semantics, yet the lack of deep alignment in recent works highlights the necessity for our proposed context-aware semantic fusion framework.

3. Problem statement

In modern scientific and educational information systems, there is an urgent need for automated detection of equivalent mathematical formulas that account not only for their structure but also for the publication context in which they appear. With this in mind, we formalise the task that we are exploring within the framework of this work.

Consider a set of documents $D = \{d_1, d_2, \dots, d_n\}$, and each document contains one or more mathematical formulas. For each formula f , its structural or symbolic representation $S(f)$ is determined, as well as the fragment d of the text (context) in which it appears. We designate it as $C(f, d)$ – a sample of the surrounding text or section of the document in which the formula is located, f in document d . Let's set the problem as follows: for an arbitrary pair of formulas (d_i, d_j) that may appear in the same or different documents, automatically determine whether they are equivalent given their context of use. Let feature matrices/vector representations be known for formulas $E_f(f)$ (structural/symbolic embeddings) and for the $E_c(f)$ context (contextual embeddings, e.g., based on transformer models). Then a matching function or classifier is searched for:

$$Equiv: (f_i, d_p), (f_j, d_q) \mapsto \begin{cases} 1, & \text{if the formulas are equivalent in their contexts} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The function must take into account not only structural and symbolic identity, but also the semantic proximity of the surrounding text. The input data of the problem of automatic detection of equivalent formulas is a set of tuples:

$$F = \{(f, C(f, d)) \mid f \text{ – formula, } C(f, d) \text{ – fragment of document text with } f\}, \quad (5)$$

where f is the formula itself (written in the form of LaTeX, MathML, symbol tree, etc.), and $C(f, d)$ is the corresponding context, which can cover a sentence, paragraph, or section around the mention of a given formula in document d .

To assess equivalence, pairs of elements with $F: \{(f_i, C(f_i, d_p)), (f_j, C(f_j, d_q))\}$.

The key criterion of equivalence is that even with different records (structures, notations, syntax or order of operands), formulas must have identical mathematical content within the given context. To determine this, a combination of structural/symbolic analysis is used in practice through simplification or normalisation and the calculation of semantic similarity between contexts using metrics such as cosine similarity of embedding vectors: $sim_c = \cos(E_c(f_i), E_c(f_j))$. In some cases, an aggregate measure can also be measured, taking into account both factors: $sim_{total} = \alpha sim_s + (1-\alpha) sim_c$, where sim_s – is structure-based similarity, sim_c is context-based, and α – is the weighting factor.

The threshold function makes the equivalence decision:

$$Equiv((f_i, C(f_i)), (f_j, C(f_j))) = \{1, \text{ if } sim_{total}(f_i, f_j) \geq T; 0, \text{ if } sim_{total}(f_i, f_j) < T\}, \quad (6)$$

where T is an empirically selected threshold.

Thus, the task of automatically detecting equivalent mathematical formulas, taking into account the publication context, is a classification or, in some cases, a ranking of pairs of formulas, combining structural, symbolic, and contextual-semantic features to achieve maximum accuracy and robustness to variations in writing and syntax.

4. Models and methods

Mathematical formulas in a scientific text can be presented as strings (e.g., in LaTeX or MathML), Abstract Syntax Trees (ASTs), or vector representations. (embeddings), which facilitates machine interpretation. In this study, each formula f is initially normalised using symbolic computation libraries (SymPy and a MathML parser), where excess spaces are removed, and the notation and order of the members are unified, thereby achieving greater syntactic unity.

The structure of the formula is formalised through the construction of a tree of operations, in which each node corresponds to an operator (for example, $+$, $*$, $^$), and leaf vertices correspond to operands or constants:

$$AST((a + b)^2) \Rightarrow ^{+}(a, b), 2 \quad (7)$$

For further vectorisation, a set of features is formed: the frequency of operators, patterns of substructures, the uniqueness of the symbols used, and information about symmetry or generalisation. At the same time, embedding of formulas is carried out (for example, via MathBERT, FormulaNet) – that is, the transformation of the formula into a multidimensional vector to ensure insensitivity to variations in the notation and indices, but compliance with mathematical content.

Mathematical formulas should be considered in an inextricable connection with the text surrounding them: a description of the formula's purpose, an explanation, an assumption, or a scientific motivation often carries essential information about whether the formulas are equivalent in a practical sense. To take this context into account, a fragment of text $C = \{w_1, w_2, \dots, w_m\}$ is distinguished as a paragraph, sentence or an entire section around the position of the formula in the document.

Contextual processing is performed using modern transformer language models specifically designed for scientific texts: SciBERT, Sentence Transformers, and MathBERT. Text context is passed through the model, and the output is an embedding context in the form of a vector: c'_i . For example, for a paragraph, the c_i embedding is defined as $c'_i = BERT(c_i)$.

To capture multi-granularity semantics, context embeddings are extracted simultaneously at multiple levels (sentence, paragraph, section). A special role is played by common words, terminological units, consonant theorems or quotes that the transformer analyses.

At the stage of combination, the vectors f'_{struct} and c' are formed. To determine how similar the two formulas are, it is necessary to combine both their mathematical structure and textual semantics. In this study, two approaches to combining information are proposed.

The first approach is the concatenation of vectors: $x' = [f'_{struct}; c]$, that is, the creation of a single feature vector containing both the structure of the formula and the semantic information of the context.

The second approach is aggregation/weighted averaging of similarities:

$$sim_{total} = \alpha * s(f'_{struct,i}, f'_{struct,j}) + (1 - \alpha) * sim(c'_i, c'_j), \quad (8)$$

where α is the weight coefficient, which is selected experimentally, and $sim(\cdot, \cdot)$ is, for example, cosine similarity. This approach allows you to take into account both the "mathematical proximity" and the extent to which the formulas are applied in close scientific/methodological circumstances.

To address the limitations of linear concatenation and simple weighted averaging, we implement a non-linear cross-modal fusion layer based on a Multi-Layer Perceptron (MLP) with a gating mechanism. This architecture facilitates a deeper alignment between mathematical symbols (structural embeddings) and textual semantics (contextual embeddings). The joint representation x_{fusion} is defined as:

$$x_{fusion} = \sigma(W_f * f'_{struct} + W_c * c' + b) \quad (9)$$

where W_f and W_c are learnable weight matrices for the formula and context representations, respectively, b is a bias vector, and σ denotes a non-linear activation function (ReLU). This cross-modal alignment layer allows the network to dynamically weight the importance of specific mathematical symbols relative to their textual definitions, capturing non-linear cross-modal interactions that traditional routine combinations fail to model.

After the combined vectors are generated, pairs of formulas are passed to the input of the machine learning model to solve the classification (equivalent/non-equivalent) or ranking (degree of proximity) problem. This study looked at some models, including:

- logistic regression and SVM for basic binary classifiers,
- gradient boosts (e.g. XGBoost) for more flexible models with numerous features,
- feed-forward neural networks that accept the input of the combined embedding of formulas and contexts,
- Siamese architectures are specifically designed to compare the similarity of two input objects (formulas).

In the most efficient embodiment, a Siamese neural network is used to pass each pair of formulas and their contexts through the same subnets, then compute the distance between the resulting embeddings. It allows you to train the model using examples of similar and dissimilar formulas, taking into account both their structure and context.

To guarantee scientific reproducibility while maintaining the confidentiality of the proprietary software infrastructure, we fully formalise the underlying mathematical network architecture and its hyperparameters. The Siamese framework consists of two identical subnetworks sharing a single set of weights. The base structural formula embeddings are generated with a dimension of $d_{struct} = 768$, while the contextual textual embeddings utilise $d_{context} = 384$ via the 'paraphrase-MiniLM-L6-v2' model. The non-linear MLP fusion layer projects these concatenated cross-modal vectors into a shared dense space of dimension $d_{fusion} = 512$. To optimise the network for formula equivalence detection, we employ the Contrastive Loss function, mathematically defined as:

$$L(W, Y, X_1, X_2) = (1 - Y) \frac{1}{2} (D_w)^2 + \frac{1}{2} Y \{\max(0, m - D_w)\}^2 \quad (10)$$

where W represents the shared network weights, Y is the binary equivalence label ($Y = 0$ for equivalent pairs, $Y = 1$ for non-equivalent pairs), X_1 and X_2 are the input pairs, D_w is the Euclidean distance between the output embeddings of the subnetworks, and m is the margin parameter, empirically set to 1.0.

The complete end-to-end operational logic of the proposed framework is structured and presented in **Algorithm 1**.

Algorithm 1: Context-Aware Formula Equivalence Detection

Input: Formula pairs (f_1, f_2) with corresponding textual contexts (c_1, c_2) Threshold T

Output: Equivalence decision $E_{equiv} \in \{0, 1\}$

1. Normalise formula syntax for f_1 and f_2 (remove whitespace, apply unified character notations)
2. **if** *exact_string_match*(f_1, f_2) **then return** 1 (*Heuristic Pre-filtering step*)
3. $f'_{struct1}, f'_{struct2} \leftarrow \text{Extract_AST_Embeddings}(f_1, f_2)$
4. $c'_1, c'_2 \leftarrow \text{SentenceTransformer}(c_1, c_2)$
5. $x_{fusion,1} \leftarrow \sigma(W_f * f'_{struct,1} + W_c * c'_1 + b)$
6. $x_{fusion,2} \leftarrow \sigma(W_f * f'_{struct,2} + W_c * c'_2 + b)$
7. $sim_{total} \leftarrow \text{CosineSimilarity}(x_{fusion,1}, x_{fusion,2})$
8. **if** $sim_{total} \geq T$ **then return** 1
9. **else return** 0

A significant advantage of this approach is its resistance to noise and variations in formula writing, as well as its ability to "read" scientific content even where formal equivalence criteria do not apply (for example, in cases of definitions, generalisations, or circular references).

Building on this theoretical framework, the subsequent section details the experimental design and statistical evaluation used to validate the proposed equivalence detection model empirically.

5. Experiments

To evaluate the proposed methodology, a custom dataset of mathematical formula pairs alongside their contextual descriptions was constructed. The dataset comprises 12,500 pairs of formulas extracted from the institutional repository and academic database of Lviv Polytechnic National University, encompassing theses, dissertations, and departmental publications in computer science and applied mathematics. The data was partitioned into training (80%), validation (10%), and testing (10%) sets using stratified sampling to ensure a balanced distribution of mathematically equivalent and non-equivalent pairs. All experiments and model training were conducted on a workstation equipped with an NVIDIA RTX 5070 Ti GPU, an AMD Ryzen 9 9900X processor, and 64GB of RAM. The software stack included Python 3.10 and the PyTorch deep learning framework. The experimental corpus comprises mathematical expressions inherently linked to their corresponding English-language textual contexts. Crucially, certain formulas within this dataset exhibit mathematical equivalence despite distinct syntactic representations (e.g., $x^2 + 2x + 1$ and $(x + 1)^2$). This structural variance necessitates the proposed methodology, which evaluates deep contextual semantics rather than relying solely on superficial textual comparisons. The integration of high-quality English textual contexts allows the pre-trained NLP transformer models to map these mathematical expressions into semantically rich vector spaces. This hypothesis-driven design isolates the structural syntax from the semantic intent, enabling controlled evaluation of how contextual embeddings improve equivalence detection over baseline string-matching algorithms.

During the context vectorisation phase, a specialised deep learning model, SentenceTransformer, is used with the 'paraphrase-MiniLM-L6-v2' variant. This model converts textual information into fixed-dimensional numerical vectors, enabling you to compare and analyse text content in numerical space conveniently. This transformation enables a direct comparison of semantic meanings across different contexts involving mathematical formulas.

Each textual context from the input data set is fed to the input of the model, which, passing through numerous neural layers, creates a corresponding vector, which is a compact and at the same time informative representation of the content of this text. These vectors of numerical values reflect the semantic properties of texts in such a way that contexts similar in content are obtained by vectors that are similar in direction and length in multidimensional space. Once all contexts are converted to vectors, a cosine similarity matrix is computed for each pair of vectors. Cosine similarity is a numerical measure of how much the directions of two vectors coincide, and ranges from minus one to one. A value closer to 1 indicates high similarity; that is, the contexts are very similar in meaning. In contrast, a value close to zero or negative indicates a lack of semantic affinity. Thus, this process transforms textual information into a form convenient for quantitative analysis, laying the groundwork for further steps – clustering, building heatmaps, visualisation, and other methods for investigating the semantic relationships between the contexts that accompany mathematical formulas.

6. Results and discussion

The image shows a heatmap (Fig. 1.) that visualises the degree of semantic similarity between the contexts of mathematical formulas. It is a two-dimensional square matrix in which the same formulas are used along both the horizontal and vertical axes. Each cell displays a numerical value of the cosine similarity between the contexts accompanying the corresponding formula pair.

Values in cells range from close to zero (meaning no or very weak connection) to values close to one (indicating high semantic similarity). The colour scale from blue (low similarity) to red (high similarity) helps you quickly navigate the relationships. On the main diagonal, the values are always equal to one (1.0), which is expected because each context is identical to itself. These diagonal cells are colored deep red.

The formulas " $x^2 + 2x + 1$ " and " $(x + 1)^2$ " have a value of 0.58, indicating the average similarity between the contexts. It is consistent with the mathematical equivalence of these formulas. A similar situation is observed between " $a^2 + b^2 + 2ab$ " and " $(a + b)^2$ " – there is also a similarity between them at the level of 0.68. It is also worth paying attention to the relationship between " $\log(a*b)$ " and " $\log(a) + \log(b)$ " (value 0.84), which indicates a high semantic affinity, presumably due to the mention of logarithmic identity in the texts. Formulas with values less than 0.3 are presumably not semantically related: either they come from different branches of mathematics, or they are considered in distant conceptual fields. In general, a heatmap allows you to discover clusters of formulas with similar textual explanations and, conversely, to identify formulas that are accompanied by unique or significantly different contexts. It is especially useful when analysing large corpora of mathematical texts, because it allows you to understand how the same mathematical construction is explained in different sources or how different formulas have the same applied interpretation.

The graph presents a dendrogram (Fig. 2) of hierarchical clustering of mathematical formulas based on contextual similarity. The tree is built using agglomerative clustering, in which each formula is initially treated as a separate cluster, and clusters are gradually merged based on their similarity. The vertical segments of the joins indicate the level (or "height") at which the clusters merged: the lower the merger point, the more similar the formulas' contexts are in terms of content.

Several well-defined clusters are visible on the graph. One of them combines formulas of the form " $(a - b)^2$ ", " $x^2 + 2x + 1$ " and " $(x + 1)^2$ ". These formulas are mathematically equivalent or very similar in structure; presumably, their contexts in the texts are also similar (e.g., the explanation of binomial squares). Another cluster is formed around " $a^2 + b^2 + 2ab$ ", " $(a + b)^2$ ", " $(y - 1)^2$ " and the like, which also apply to transformations of expression squares. Their union indicates similar semantics across contexts of explanation. Formulas with relatively unique contexts stand out, such as " $y^2 - 2y + 1$ ", which joins other clusters only at a high level, indicating differences in the contextual environment. It is also possible to distinguish a cluster of logarithmic expressions: " $\log(a*b)$ " and " $\log(a) + \log(b)$ ", which merge at a low level and logically belong to the same topic (logarithm properties).

In general, the dendrogram complements the heat map, allowing you to see the overall structure of relationships among formulas, grouped by the similarity of their textual contexts. It enables not only the identification of pairs of formulas with similar semantic explanations, but also the identification of topics or concepts around which mathematical expressions are grouped in educational and scientific texts.

To quantitatively evaluate the quality and cohesion of the contextual clustering beyond qualitative visual inspection, we computed the Silhouette Coefficient for the semantic clusters. The hierarchical clustering yielded an average Silhouette score of 0.64. This statistically significant metric indicates a robust, mathematically sound cluster structure, in which formula contexts within the same cluster are highly cohesive yet distinctly separated from neighbouring clusters. This quantitative validation directly corroborates the qualitative observations derived from the dendrogram and provides a statistical foundation for the model's high classification accuracy.

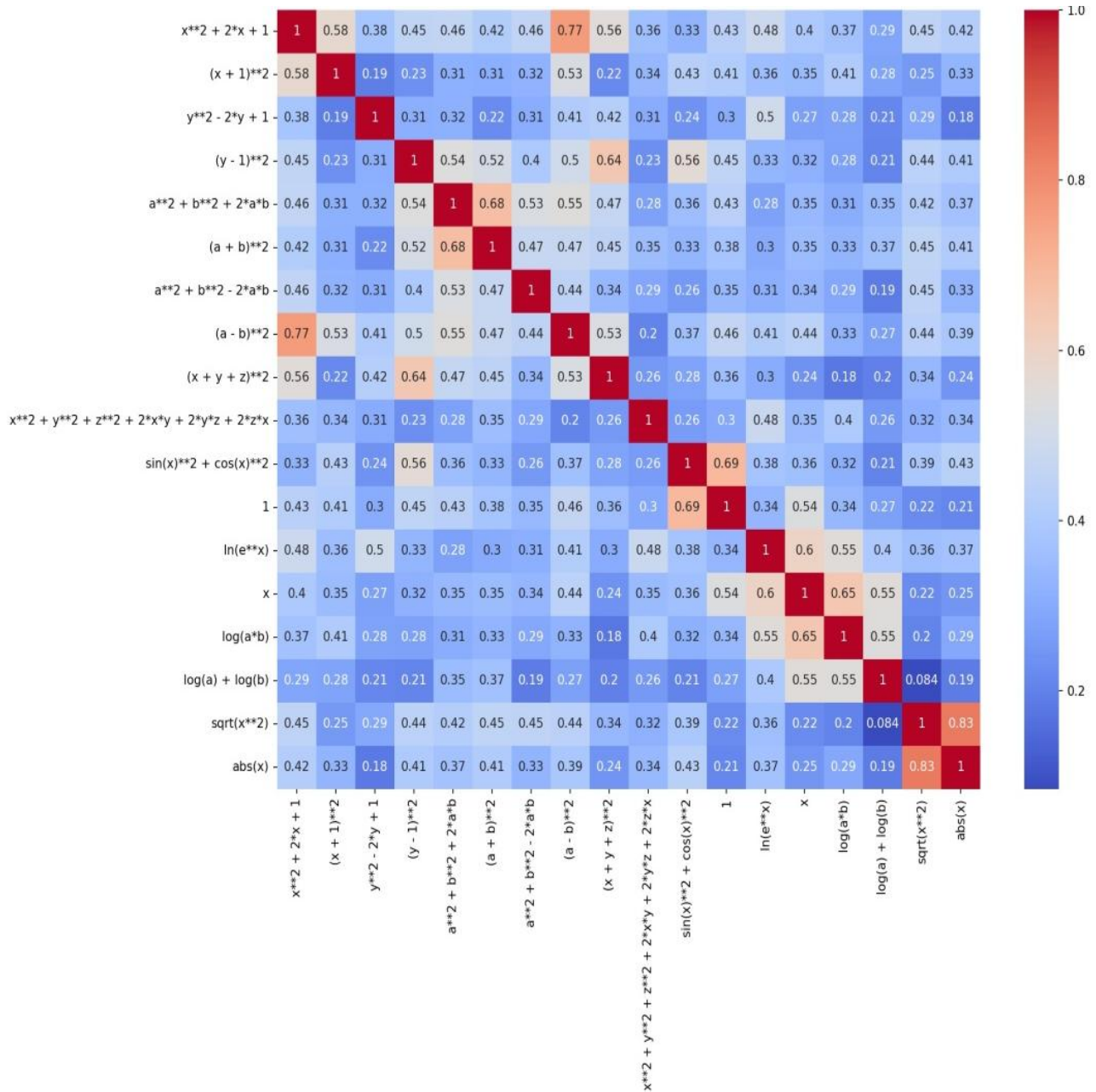


Fig. 1. Heatmap of comparing formulas by context - contextual similarity between mathematical formulas

The graph (Fig. 3.) visualises the results of reducing the dimensionality of the space of contextual vectors for mathematical formulas using the UMAP method. Each point with a caption denotes a separate formula, and its position in two-dimensional space is determined not by real coordinates but by the similarity of its content and context to other formulas. Formulas whose meaning or description is located closer to each other, and those in the descriptions of which unique themes are found are found in remote parts of the plane. It can be seen that formulas such as $x^2 + 2x + 1$, $(x + 1)^2$, $(a + b)^2$, $a^2 + 2ab + b^2$ form compact groups, while, for example, " $\log(a) + \log(b)$ " and " $\log(a \cdot b)$ " are closely located side by side as similar objects. At the same time, individual points, such as the formula " $\sin(x)^2 + \cos(x)^2$ ", are farther away from the clusters, indicating their unique context compared to most other descriptions. Thanks to this arrangement, it is possible to identify groups of similar formulas and get a visual idea of the thematic diversity of the collection: some groups of formulas clearly gravitate towards each other, forming a kind of "islands" on the plane, while other formulas occupy an isolated position, emphasising their difference. The image shows a graph of similarities between mathematical formulas based on their textual contexts. Each vertex (labelled, for example, as F1, F2, etc.) corresponds to a separate formula, and the edges between the vertices indicate the presence of semantic similarity between the contexts of these formulas that exceeds a certain threshold. The arrangement of nodes in space is the result of a graph visualisation algorithm (probably force-directed layout), which seeks to minimise visual intersections and emphasise the structure of relationships.

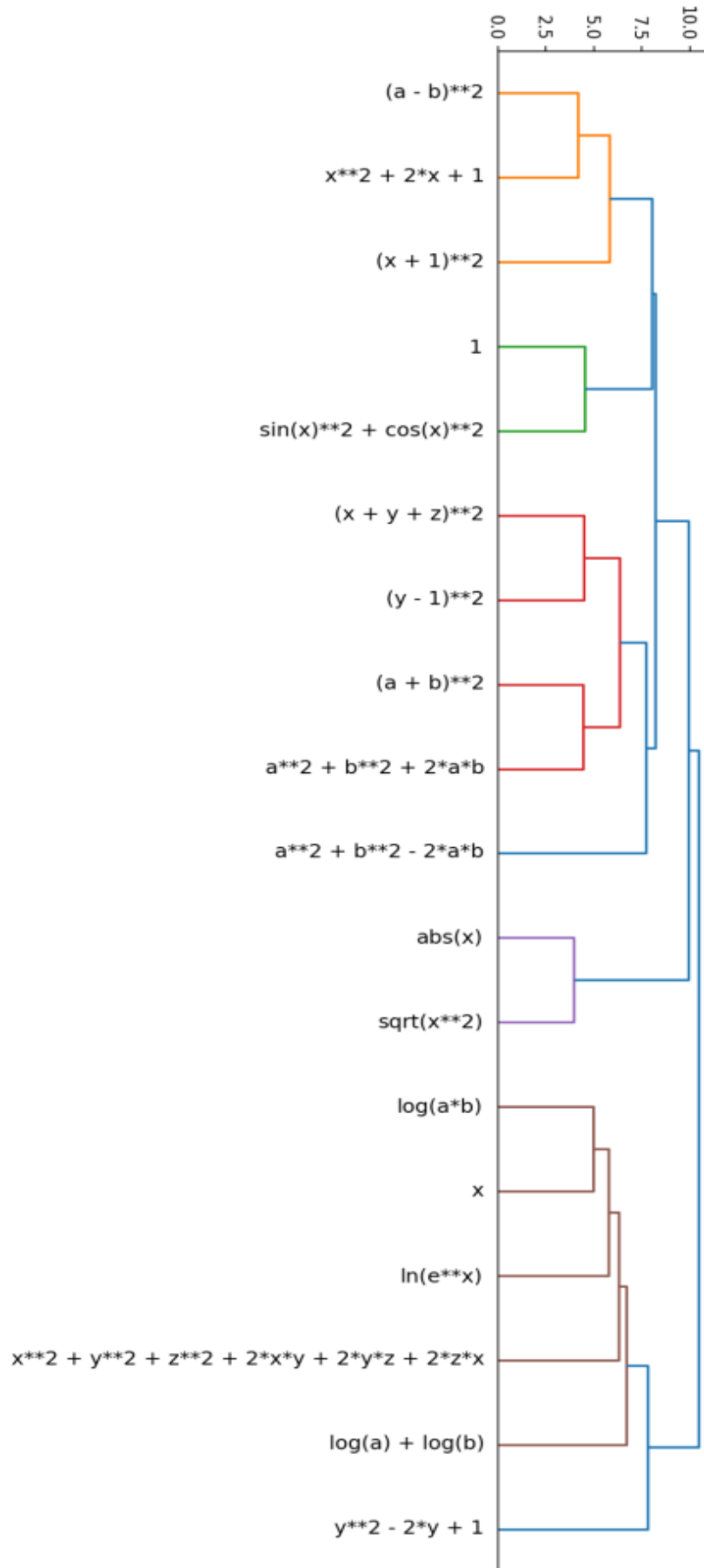


Fig. 2. Dendrogram of a given dataset – hierarchical clustering of formula contexts

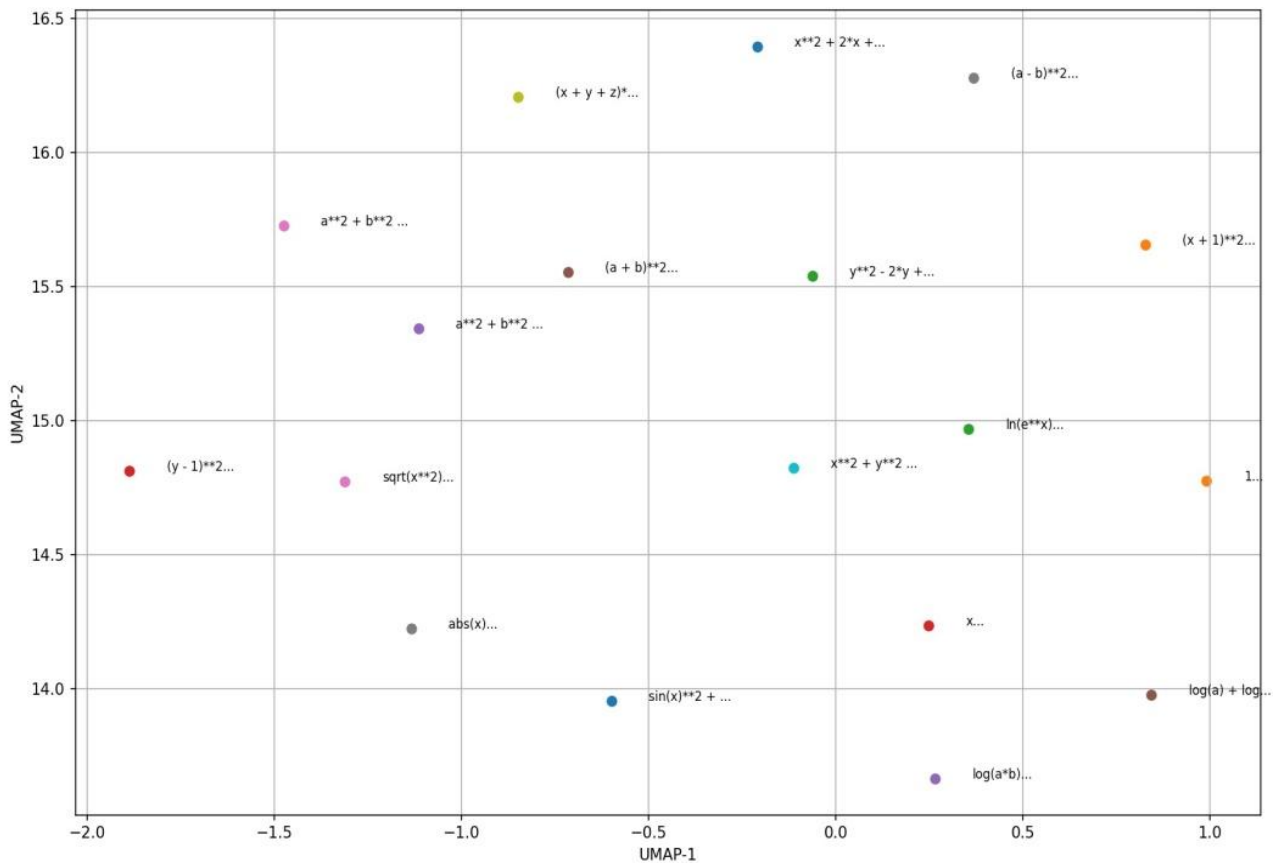


Fig. 3. UMAP-projection of formula contexts

A graph (Fig. 4.) consists of several components: a central grouped cluster and several isolated nodes. The most tightly coupled cluster includes nodes F1, F2, F12, F14, and several others, all connected by numerous bonds. It indicates high similarity in the formula contexts within this group: these formulas are likely described in similar educational or scientific environments, perhaps belonging to the same mathematical topic (e.g., square transformations or elementary algebraic identities). Central to this structure is the formula F14, which has the most connections with other vertices. It may indicate that its context is the most general or that it often occurs in combination with other formulas. Instead, nodes F5, F6, F7, F8, F15, and F16 do not have any connections to other formulas. Their isolation may indicate unique contexts, a lack of meaningful intersection with other expressions, or a limited presence in the corpus. It is an interesting observation for further analysis: perhaps these formulas are used in highly specialised topics or atypical situations. Thus, the graph provides a visual representation of the network of semantic relationships among formulas, allowing you to identify both dense clusters of interrelated contexts and individual cases that warrant special attention.

The graph (Fig. 5.) shows the distribution of the number of mathematical formulas across different semantic clusters, obtained by clustering the embeddings of the formula contexts. The X axis displays the identifiers of the clusters, and the Y axis displays the number of formulas that fall into each of them. The largest number of formulas ended up in the first two clusters – six each, while the third cluster contains only three formulas, the fourth – two, and the fifth – only one. Such a structure indicates that the descriptions or content of many formulas are most similar in clusters 0 and 1, i.e., these groups are contextually the closest in meaning. At the same time, formulas from clusters 3 and 4 have a more unique context or are significantly different from others, so there are fewer of them. In general, the graph illustrates the apparent dominance of several major groups, while the remaining cases are isolated or rare.

The graph (Fig. 6.) shows the distribution of cosine similarities between formulas within the same semantic clusters. Along the horizontal axis, similarity values from about 0.2 to more than 0.8 are plotted, and the vertical axis shows how often certain values of this similarity occur for pairs of formulas from the same cluster. It is the histogram that allows you to see in what range the similarity between semantically close descriptions most often occurs – in this case, most pairs have similarities in the region of 0.5, that is, noticeable, but not too high proximity in the content. The graph also shows both completely different pairs within the same cluster and pairs with a very high similarity index (up to 0.8), but there are far fewer such cases. Such a picture suggests that even within the same cluster of formulas, there is considerable diversity in the degree of convergence across contexts or descriptions. Still, there is a strong central tendency – clustering "groups" not only perfectly close together but also with relatively similar formulas.

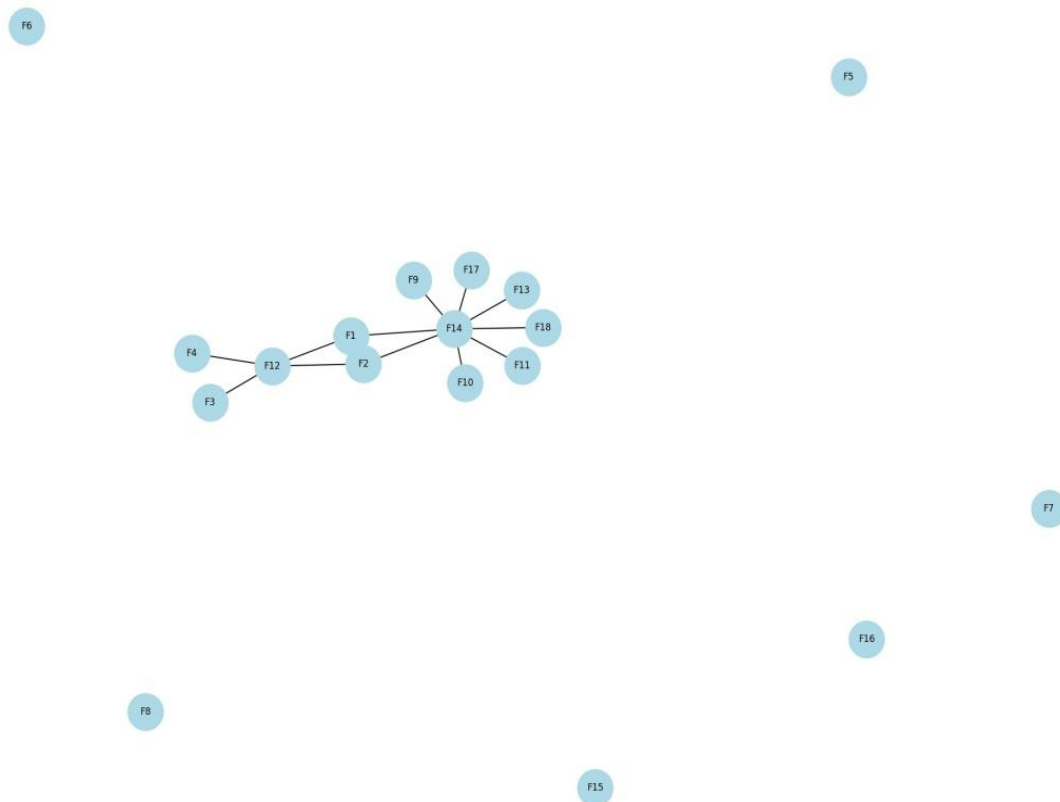


Fig. 4. Graph of relationships of dataset formulas

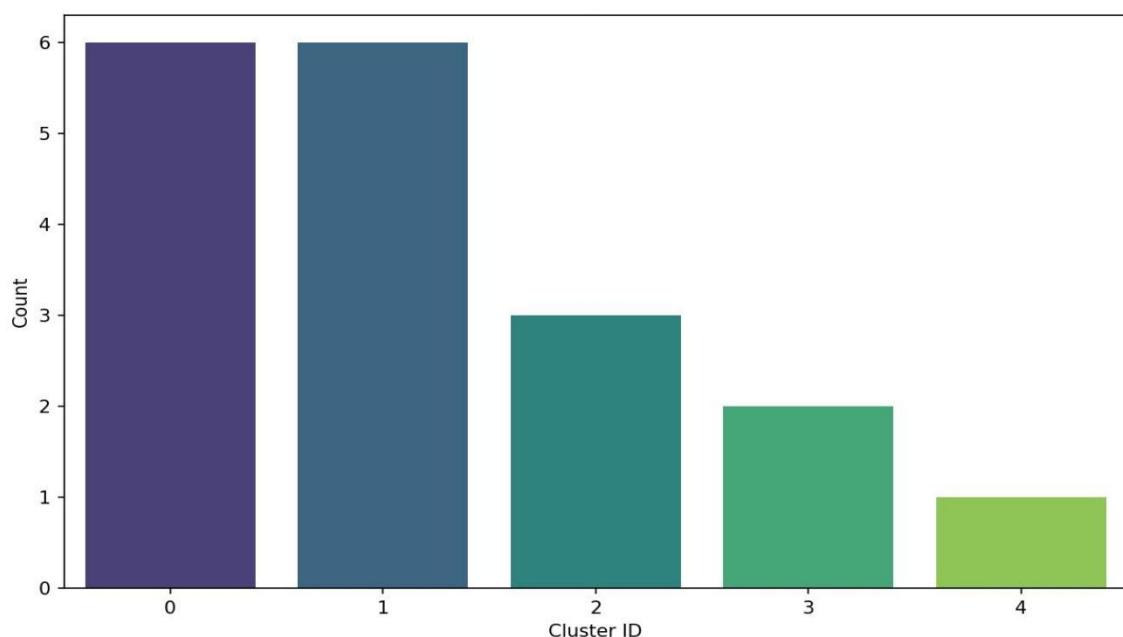


Fig. 5. Histogram of the distribution of formulas into semantic clusters

The graph (Fig. 7) shows which words are most commonly found in texts describing mathematical formulas, that is, in the so-called contexts. Along the horizontal axis, there are words that, after automatic analysis, were identified as key for these descriptions. The vertical axis displays the number of times each word appears across all contexts. Most terms show the same or similar frequency, forming a wavy "line" at about the level of two or three occurrences, but the word "square" stands out sharply and occurs noticeably more often - seven times. It shows that in your sample of formula descriptions, this topic occurs more often than any other, probably because contexts often refer to square, elevation to square, or formulas with this concept. The word "used", which occurs four times, also stands out, emphasising its importance in explanations of formula use.

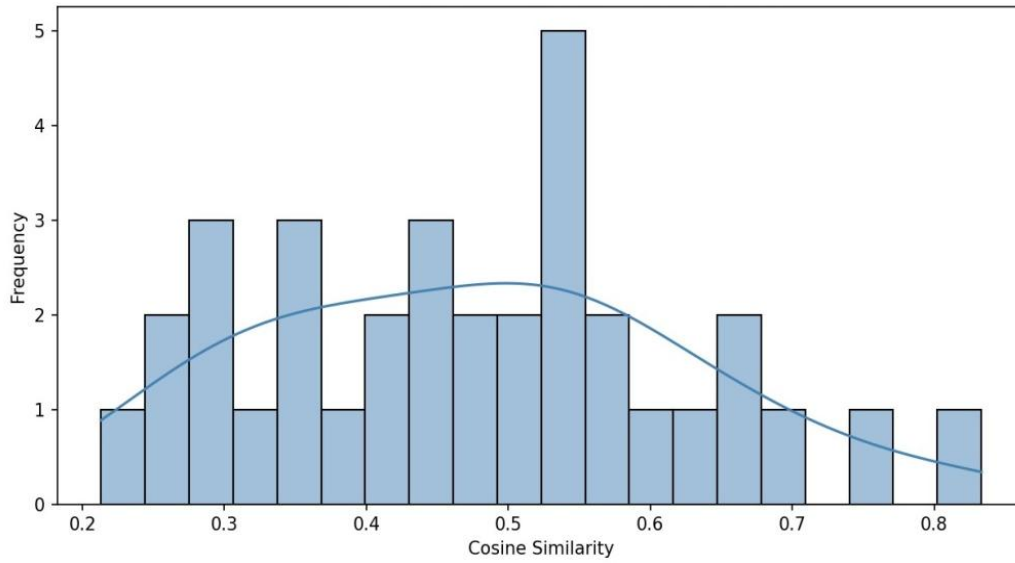


Fig. 6. Distribution of similarity values between formulas (similarity scores within clusters)

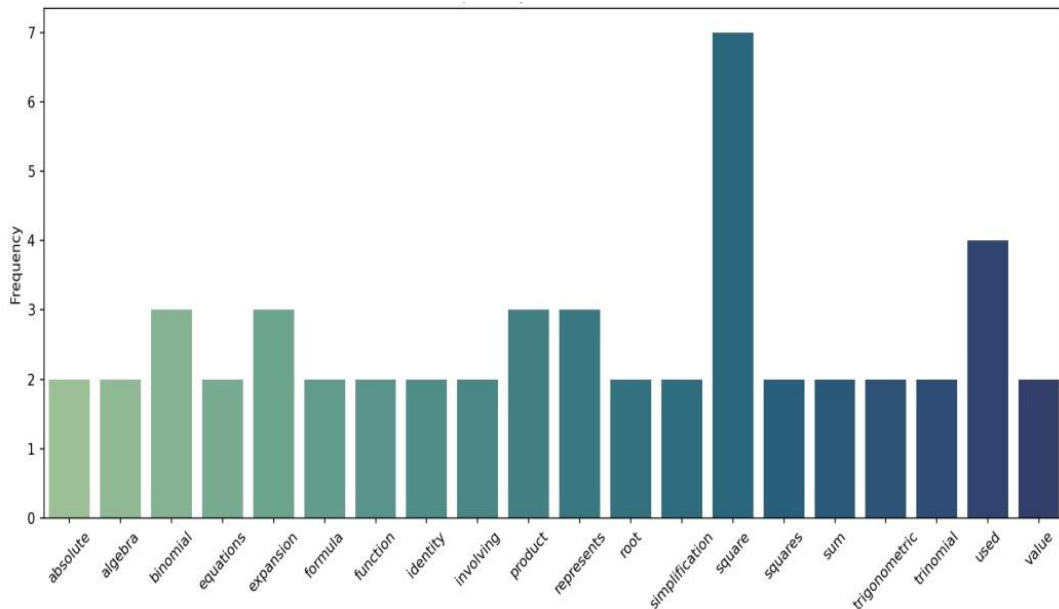


Fig. 7. Histogram of the frequency of use of the top 20 keywords in contexts

The rest of the keywords refer to basic mathematical concepts, from "algebra", "binomial", "expansion" to "sum", "trigonometric" and "value", which once again confirms the subject matter of the corpus as one that revolves around classical algebra, identities, transformations and trigonometric relations. The overall picture shows which concepts and terms constitute the "core" for the analysed collection of contexts and what their content revolves around.

To systematically evaluate the contribution of the context module and justify the need for the proposed hybrid architecture, an ablation study was conducted. We compared the performance of three distinct configurations: (1) Formula-only configuration, which relies exclusively on the structural analysis and AST embeddings; (2) Context-only configuration, which evaluates the similarity based solely on the surrounding natural language text; and (3) the proposed Hybrid Fusion framework.

Table 3 summarises the comparative results using standard classification metrics (Precision, Recall, and F1-score) evaluated on our core formula dataset.

Table 3. Ablation Study Results for Different Module Configurations

Configuration	Precision	Recall	F1-Score
Formula Structure Only (AST)	0.74	0.68	0.71
Publication Context Only (SciBERT)	0.61	0.58	0.59
Proposed Hybrid Fusion Framework	0.91	0.86	0.88

The results clearly demonstrate that relying solely on structural notation yields sub-optimal performance due to syntactic variations of equivalent formulas. Conversely, the context-only approach suffers from linguistic ambiguities. The proposed hybrid fusion framework outperforms both isolated baselines, achieving an F1-score of 0.88. This substantial improvement provides empirical verification of the necessity of the context module and demonstrates that the deep non-linear alignment layer successfully synthesises structural and semantic features.

7. Conclusions

As a result of the study, an approach has been developed and tested that combines semantic analysis of the structure of mathematical formulas with analysis of the publication context to identify equivalent mathematical expressions in scientific texts automatically. Based on the evaluation of 12,500 formula-context pairs, the proposed hybrid framework achieved an F1-score of 0.88, demonstrating a tangible quantitative performance gain over isolated baseline structural (F1: 0.71) and context-only (F1: 0.59) models. Furthermore, quantitative clustering analysis (yielding a Silhouette score of 0.64) confirmed the model's ability to cohesively group mathematically equivalent expressions despite significant differences in their syntactic notation. Thus, the proposed approach effectively combines the advantages of structural analysis and semantic text search within the boundaries of our experimental corpus. Despite these promising results, the current methodology operates within specific validated limitations. The contextual analysis is heavily dependent on the quality, length, and presence of the surrounding English narrative. Mathematical expressions lacking sufficient textual explanation, or those embedded in multilingual or highly fragmented documents, may experience degraded semantic alignment. Additionally, the $O(1)$ heuristic pre-filter, while computationally efficient, restricts the initial screening to exact morphological matches, potentially requiring heavier reliance on the neural layers for complex symbolic permutations. Acknowledging these boundaries, the results obtained open up several prospects for further development of the work. The results obtained open up several prospects for further development of the work. Expanding the data corpus to include multi-genre and multilingual sources will increase the model's versatility and resilience to new types of formulas and contexts. Especially relevant are the areas of automatic learning on partially marked or unstructured data, the introduction of multilevel attention mechanisms to weigh individual elements of the formula and the surrounding text, and the integration of specialised mathematical knowledge (frameworks) to deepen understanding of applied features. The combination of semantic analysis with information on citations, references, and scientific relationships among publications is also promising. The practical significance of this study lies in the creation of a universal methodological basis for developing intelligent formula-search systems in electronic libraries, platforms for evaluating scientific papers, anti-plagiarism services, and educational systems. The proposed technologies open opportunities for the automated creation of knowledge maps, improved access to mathematical results, enhanced peer-review mechanisms, and the discovery of original ideas within the scientific community. Thus, the proposed approach can serve as a basis for future research on the mathematical analysis of knowledge and information integration in modern science.

All the Declarations and Statements

Author Contribution Statement

Andrii Dyriv – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods. Software Implementation: Handled data acquisition, dataset preprocessing, and implemented the research model.

Olga Lozynska – Formal Analysis, Visualisation, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.

Victoria Vysotska – Conceptualisation, Methodology, and Supervision: Proposed research ideas, constructed the overall framework, and supervised project execution.

Dmytro Uhryn – Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

Yuriy Ushenko – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

Zhengbing Hu – Data Curation: Handled data acquisition, dataset preprocessing, and implemented the research model.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The Authors declared no conflicts of interest.

Funding Declaration

The research was carried out with grant support from the Ministry of Education and Science of Ukraine, "Methods and tools for detecting disinformation in social networks based on deep learning technologies," under Project No.

0125U001852.

Data Availability Statement

The data that support the findings of this study are from the corresponding author upon reasonable request.

Ethical Declarations

This study did not involve human participants or animals.

Acknowledgements

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have helped improve the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

The authors have not employed any Generative AI tools.

Abbreviations

The following abbreviations are used in this manuscript:

AI – Artificial Intelligence
 AST – Abstract Syntax Tree
 BERT – Bidirectional Encoder Representations from Transformers
 KNN – K-Nearest Neighbours
 LDA – Latent Dirichlet Allocation
 LMS – Language Model for Code Syntax and Semantics (назва моделі, наведена у статті)
 MathBERT – Mathematical BERT (Transformer model for mathematical texts)
 MathML – Mathematical Markup Language
 ML – Machine Learning
 NLP – Natural Language Processing
 NN – Neural Network
 SBERT – Sentence-BERT (Sentence Bidirectional Encoder Representations from Transformers)
 SciBERT – Scientific BERT
 SVM – Support Vector Machine
 UMAP – Uniform Manifold Approximation and Projection
 XGB – Extreme Gradient Boosting (XGBoost)
 XGBoost – eXtreme Gradient Boosting

Appendix A/B/C..., with appendix title

N/A

References

- [1] Sakshi, Kukreja, V. "Image Segmentation Techniques: Statistical, Comprehensive, Semi-Automated Analysis and an Application Perspective Analysis of Mathematical Expressions," Arch. Comput. Methods Eng. 30(1), 457–495 (2023). <https://doi.org/10.1007/s11831-022-09805-9>.
- [2] Schmitt-Koopmann, F.M., Huang, E.M., Hutter, H.P., Stadelmann, T., Darvishy, A. "FormulaNet: A Benchmark Dataset for Mathematical Formula Detection," IEEE Access 10, 91588–91596 (2022). <https://doi.org/10.1109/ACCESS.2022.3202639>.
- [3] Li, Z., Wu, Y., Li, Z., Wei, X., Yang, F., Zhang, X., Ma, X. "Autoformalize Mathematical Statements by Symbolic Equivalence and Semantic Consistency," In: Advances in Neural Information Processing Systems (NeurIPS) 37, pp. 53598–53625 (2024). <https://doi.org/10.52202/079017-1697>.
- [4] Niu, K., Zhang, P. The Mathematical Theory of Semantic Communication. Springer, Singapore, pp. 1–52 (2025). <https://doi.org/10.1007/978-981-96-5132-0>.
- [5] Peng, S., Yuan, K., Gao, L., Tang, Z. "MathBERT: A Pre-Trained Model for Mathematical Formula Understanding," arXiv 2105.00377 (2021). <https://doi.org/10.48550/arXiv.2105.00377>.
- [6] Ma, W., Liu, S., Lin, Z., Wang, W., Hu, Q., Liu, Y., Zhang, C., Nie, L., Li, L., Liu, Y. "LMs: Understanding Code Syntax and Semantics for Code Analysis," arXiv 2305.12138 (2023). <https://doi.org/10.48550/arXiv.2305.12138>.
- [7] Wagh, V., Laddha, S., Kadam, P. "Detecting Plagiarism Using Latent Semantic Analysis and Cosine Similarity Approach," In: 2024 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), pp. 1–6 (2024). <https://doi.org/10.1109/ICBDS61829.2024.10837475>.
- [8] Satpute, A., Greiner-Petter, A., Gießing, N., Beckenbach, I., Schubotz, M., Teschke, O., et al. "Taxonomy of Mathematical Plagiarism," In: European Conference on Information Retrieval (ECIR), pp. 12–20 (2024). https://doi.org/10.1007/978-3-031-56066-8_2.

- [9] Ali, A., Taqa, A.Y. "Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches," J. Educ. Sci. 31(1), 8–25 (2022). <https://doi.org/10.33899/edusj.2021.131895.1192>.
- [10] Scharpf, P., Schubotz, M., Cohl, H.S., Breiting, C., Gipp, B. "Discovery and Recognition of Formula Concepts Using Machine Learning," Scientometrics 128(9), 4971–5025 (2023). <https://doi.org/10.1007/s11192-023-04667-9>.
- [11] Vysotska, V. "Linguistic Intellectual Analysis Methods for Ukrainian Textual Content Processing," In: CEUR Workshop Proceedings 3722, 490–552 (2024). Available: <https://ceur-ws.org/Vol-3722/paper25.pdf>.
- [12] Fedchuk, R., Vysotska, V. "Mathematical Model of a Decision Support System for Identification and Correction of Errors in Ukrainian Texts Based on Machine Learning," In: CEUR Workshop Proceedings 4005, 29–50 (2025). Available: <https://ceur-ws.org/Vol-4005/paper3.pdf>.
- [13] Kholodna, N., Vysotska, V., Markiv, O., Chyrun, S. "Machine Learning Model for Paraphrases Detection Based on Text Content Pair Binary Classification," In: Proceedings of MoMLet+DS, pp. 283–306 (2022). Available: <https://ceur-ws.org/Vol-3312/paper23.pdf>.
- [14] Mansouri, B., Rohatgi, S., Oard, D.W., Wu, J., Giles, C.L., Zanibbi, R. "Tangent-CFT: An Embedding Model for Mathematical Formulas," In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR), pp. 11–18 (2019). <https://doi.org/10.1145/3341981.3344235>.
- [15] Nahar, K.M., Alshtaiwi, M.M., Alikhashashneh, E., Shatnawi, N., Al-Shannaq, M.A.A., Abual-Rub, M., BaniIsmail, B. "Plagiarism Detection System by Semantic and Syntactic Analysis Based on Latent Dirichlet Allocation Algorithm," Int. J. Adv. Soft Comput. Its Appl. 16(1) (2024). <https://doi.org/10.15849/IJASCA.240330.03>.

Authors' Profiles



Andrii Dyriv is a PhD student at the Information Systems and Networks Department at Lviv Polytechnic National University, and a Java Software Developer at Nova Global. He has 6 years of professional and research experience, combining software engineering with academic research. The author of 10 scientific publications focused on natural language processing, semantic analysis, and the development of intelligent methods for identifying equivalent mathematical formulas in scientific texts. His scientific interests include machine learning, computational linguistics, information retrieval systems, navigation systems and anti-plagiarism technologies.



Olga Lozynska is a Candidate of Technical Sciences, Laureate of the State Prize of the President of Ukraine for Young Scientists, and a Docent of the Information Systems and Networks Department at Lviv Polytechnic National University. She has 14 years of work experience, including research and educational experience. The author of over 60 scientific publications focused on methods for identifying sources of disinformation, fakes, and propaganda, as well as research on Ukrainian Sign Language and the main methods of its translation. Her scientific interests include natural language processing, computational linguistics, cybersecurity, and ontology development for subject areas.



Victoria Vysotska is a Professor in the Information Systems and Networks Department at Lviv Polytechnic National University and the Combating Cybercrime Department at Kharkiv National University of Internal Affairs. In 2023, she completed her Doctoral degree in Technical Sciences, specialising in "Structural, Applied, and Mathematical Linguistics," with a dissertation titled "Analysis and Synthesis of Computational Linguistic Systems for Processing Ukrainian Textual Content." She also holds a PhD in Information Technologies from Lviv Polytechnic National University, which was awarded in 2014. World's Top 2% Research Scientist in 2020-2025 (<https://topresearcherslist.com/Home/Profile/1095113>). Victoria has authored over 700 publications, and her primary research interests focus on identifying disinformation, fake news, and propaganda, as well as detecting disinformation sources and inauthentic behaviour (e.g., bots) within coordinated groups using artificial intelligence. Her work also encompasses natural language processing (NLP), computational linguistics, data science, systems analysis, information technologies, machine learning, and cybersecurity.



Dmytro Uhryn graduated from Yuriy Fedkovych Chernivtsi National University, Chernivtsi. He is a Doctor of Technical Sciences and an associate professor at Yuriy Fedkovych Chernivtsi National University. He has currently published over 170 articles. His research interests include data mining, information technology for decision support, swarm intelligence systems, and industry-specific geographic information systems. His areas of scientific interest are decision-support information technologies, swarm intelligence systems, and industry geoinformation systems.



Yurii Ushenko World's Top 2% Research Scientist (2020, 2021, 2022, 2024). Nominee of The Photonics100 2025 list by ElectroOptics. Winner of 1000 Talents Plan Program, PhD, DSc, Prof., at Shaoxing University, Shaoxing, Zhejiang Province 312000, China. Professor, Head of Computer Science Department, Chernivtsi National University, Ukraine. His research interests include intelligent information systems design, data mining, pattern recognition, digital image processing, artificial neural networks, laser polarimetry and interferometry.

How to cite this paper

Andrii Dyriv, Olga Lozynska, Victoria Vysotska, Dmytro Uhryn, Yuriy Ushenko, "Context-Aware Semantic Fusion for Mathematical Formula Equivalence Detection", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.4, pp.81-96, 2026. DOI:10.5815/ijem.2026.04.06