UDC 81'42:004.912
DOI https://doi.org/10.32782/2307-1222.2025-60-30

METHODS OF CONCEPT EXTRACTION IN LITERARY WORKS

Nataliia KUNANETS
Doctor of Science in Social Communications, Professor,
Professor at the Department of Information Systems and Networks
Lviv Polytechnic National University
12 Bandery str., Lviv
ORCID: 0000-0003-3007-2462
nek.lviv@gmail.com

Maksym YAROMYCH
Postgraduate Student at the Department of Applied Linguistics
Lviv Polytechnic National University
12 Bandery str., Lviv
ORCID: 0009-0005-3299-6695
yamax0312@gmail.com

The article explores both traditional and contemporary approaches to concept extraction in literary texts, with a particular focus on lexical analysis, semantic parsing, ontological modeling, and automated techniques based on large language models. Traditional manual methods are valued for their ability to capture nuanced literary meanings – such as symbolism, metaphor, and intertextual references – while taking into account cultural and stylistic context. However, their reliance on intensive human interpretation makes them difficult to apply to large text corpora and comparative studies.

Modern automated approaches, especially those utilizing transformer architectures like BERT, introduce significant advantages in processing speed and scalability. Through mechanisms such as self-attention, these models effectively identify long-range contextual relationships and latent patterns within texts, enabling rapid detection and classification of key concepts across extensive datasets. Yet, the article emphasizes that these systems still face limitations when dealing with the figurative richness and semantic ambiguity inherent in literary discourse.

The practical component of the research involves an experimental analysis of the concept of tolerance in articles from major English-language media outlets, including The New York Times, BBC News, and The Guardian. Automated extraction methods demonstrated strong potential for identifying general trends and conceptual usage patterns. Nonetheless, the findings underscore the need for expert interpretation to refine outputs, especially in cases involving subtle semantic shifts or interdisciplinary cultural references.

In conclusion, the article proposes an integrative methodological framework that combines automated processing with human expertise. Automated tools are

recommended for the preliminary structuring and classification of large volumes of textual data, while expert scholars are responsible for interpretative depth and validation. Future research directions include enhancing LLM adaptation for literary texts, building specialized training corpora, and incorporating ontological models to improve conceptual precision and reliability in literary studies.

Key words: *conceptual analysis, large language models, automated concept extraction, ontological modeling, BERT, natural language processing.*

МЕТОДИ ВИДІЛЕННЯ КОНЦЕПТІВ У ЛІТЕРАТУРНИХ ТВОРАХ

Наталія КУНАНЕЦЬ
доктор наук із соціальних комунікацій, професор,
професор кафедри інформаційних систем та мереж
Національного університету «Львівська політехніка»
вул. С. Бандери, 12, м. Львів
ORCID: 0000-0003-3007-2462
nek.lviv@gmail.com

Максим ЯРОМИЧ
аспірант кафедри прикладної лінгвістики
Національного університету «Львівська політехніка»
вул. С. Бандери, 12, м. Львів
ORCID: 0009-0005-3299-6695
yamax0312@gmail.com

Стаття досліджує як традиційні, так і сучасні підходи до виокремлення концептів у літературних текстах, приділяючи особливу увагу лексичному аналізу, семантичному розбору, онтологічному моделюванню та автоматизованим технікам на основі великих мовних моделей. Традиційні ручні методи цінуються за їхню здатність охоплювати нюансові літературні значення, такі як символізм, метафора та інтертекстуальні відсилання, з урахуванням культурного й стильового контексту. Однак їхня залежність від інтенсивної людської інтерпретації робить їх складними для застосування до великих корпусів текстів і порівняльних досліджень. Сучасні автоматизовані підходи, особливо ті, що використовують архітектури трансформерів, як-от BERT, забезпечують значні переваги у швидкості обробки та масштабованості. Завдяки механізмам самоуваги ці моделі ефективно виявляють довготривалі контекстуальні зв'язки та приховані патерни в текстах, даючи змогу швидко визначати й класифікувати ключові концепти у великих наборах даних. Водночас наголошено, що такі системи досі мають обмеження під час роботи з образною насиченістю та семантичною неоднозначністю, притаманними літературному дискурсу.

Практична частина дослідження включає експериментальний аналіз концепту толерантності у публікаціях провідних англомовних медіа, зокрема The New

York Times, BBC News та The Guardian. Автоматизовані методи виокремлення продемонстрували значний потенціал у виявленні загальних тенденцій і патернів використання концептів. Проте результати підкреслюють необхідність експертної інтерпретації для уточнення отриманих даних, особливо у разі тонких семантичних зсувів або міждисциплінарних культурних відсилань.

У підсумку стаття пропонує інтегративну методологічну рамку, яка поєднує автоматизовану обробку з людською експертизою. Автоматичні інструменти рекомендується використовувати на початковому етапі для структурування та класифікації великих обсягів текстових даних, тоді як експертні дослідники відповідають за інтерпретаційну глибину та валідацію. Подальші напрями досліджень включають удосконалення адаптації LLM до літературних текстів, створення спеціалізованих навчальних корпусів та інтеграцію онтологічних моделей із метою підвищення точності та надійності концептуального аналізу в літературознавстві.

Ключові слова: *концептуальний аналіз, великі мовні моделі, автоматизоване виділення концептів, онтологічне моделювання, BERT, опрацювання природної мови.*

Problem Statement. One of the key tasks of modern linguistics and literary studies is the effective identification and analysis of key concepts in literary works. Concepts are central units of analysis that help us deeply understand the meaning, ideas, subtext, and the author's intention behind a text. Therefore, studying concepts is essential for revealing the cultural, social, and philosophical aspects of literary texts, as well as for explaining the impact of a work on its audience and its place in the literary process [8; 9]. However, traditional methods of concept identification – such as manual lexical analysis, semantic interpretation, and comparative analysis – have several serious limitations. These methods are time-consuming, depend heavily on the researcher’s subjective interpretation and expertise, and are often not suitable for analyzing large text corpora. This significantly limits the scope of comparative studies and large-scale textual analysis [9].

The problem of identifying concepts in literary texts is actively studied both in traditional philology and in the context of modern natural language processing technologies. Traditional methods mainly rely on manual work, the researcher’s expertise, and deep understanding of the text’s context and cultural background. These methods are effective for in-depth analysis, but their scale is often limited. One of their main strengths is the ability to accurately identify and explain symbolic, cultural, and author-specific meanings that cannot be fully captured by automated tools. For example, manual semantic analysis helps uncover the ambiguity and individual interpretation of concepts by the author, which is especially important when analyzing literary works where context and subtext play a central role.

With the rise of modern information technologies, large language models such as BERT, GPT, RoBERTa, and others – based on transformer architecture – have become widely used for conceptual analysis. Later developments, such as RoBERTa and

ALBERT [3], have shown the strong potential of these models in concept extraction, thanks to their ability to consider contextual and semantic nuances in texts. The automation of concept extraction is also closely linked to ontology-based modeling. Combining ontologies with large language models makes it possible to structure knowledge more effectively and improve the accuracy and relevance of conceptual analysis, especially in specialized fields such as literary studies or biomedicine [10; 11; 26]. For instance, Bartalesi and Meghini investigated the use of ontologies to structure knowledge in literary texts, using the works of Dante Alighieri as a case study [4]. Other researchers emphasize the potential of automatic ontology generation and enrichment based on natural language analysis, which further supports the relevance of combined approaches [18; 29].

At the same time, some publications highlight the challenges of automated approaches – for example, the occurrence of «hallucinations», where the model produces irrelevant or incorrect concepts. This issue has been addressed in detail in [20], where the researchers suggest that ontology-based approaches can help reduce such risks. Despite the large number of studies demonstrating the potential of large language models and their integration with ontologies in concept extraction, several important questions remain underexplored. In particular, the limits of effectiveness of automated models when working with complex literary texts are still not clearly defined. Such texts are often rich in metaphor, symbolism, and layered meanings. It is still unclear to what extent large language models can truly take into account deep cultural and contextual features, which are crucial for proper interpretation of literary works. In addition, the issue of how best to integrate automated and traditional manual methods of analysis remains open. So far, there are no clear guidelines on when automated approaches should be used and when manual, expert-based analysis is necessary to achieve the highest accuracy and completeness of results.

The rapid growth of textual data and the increasing interest in analyzing large collections of texts require more efficient, automated approaches. With the development of information technologies and natural language processing methods, new automated techniques for text analysis have emerged. These methods offer clear advantages over traditional ones. Of particular interest in this context are large language models such as BERT, GPT, RoBERTa, and T5. Thanks to their transformer-based architecture, these models can process large amounts of textual information quickly and accurately, taking into account contextual and semantic relationships between words. Still, despite their potential, the use of automated models in literary studies remains underexplored. Questions about the effectiveness and accuracy of such models in specific fields like literary analysis remain open and require deeper academic investigation. Special attention should be given to how large language models can be adapted to analyze literary texts, which often include stylistic, metaphorical, and symbolic elements [27].

Thus, there is a relevant need to compare the effectiveness of traditional and modern methods for identifying concepts in literary texts. It is important to understand the specific advantages of automated approaches compared to manual ones, determine when

automated methods are appropriate, and when expert human analysis is still necessary. Another issue is how to best integrate these two approaches to achieve the most accurate and reliable results. In light of this, the aim of this study is to conduct a comparative analysis of the effectiveness of traditional and modern automated methods for identifying concepts in literary texts, as well as to determine optimal approaches for their combined use. To achieve this aim, the research will apply the analytical method to examine existing techniques for studying concepts, the comparative method to identify the advantages and disadvantages of each approach, and the experimental method to evaluate the effectiveness of automated models in identifying the concept of "tolerance" in texts from respected media sources such as the New York Times, BBC News, and The Guardian.

Aim of the Study. The aim of this study is to conduct a comparative analysis of the effectiveness of traditional and modern automated methods for identifying concepts in literary texts, as well as to determine optimal approaches for their combined use.

Presentation of the Main Material. To achieve the aim of conducting a comparative analysis of the effectiveness of traditional and modern automated methods for identifying concepts in literary texts, the research applies a combination of analytical, comparative, and experimental methods. The analytical method is used to examine existing techniques for studying concepts, including manual lexical approaches and automated methods based on large language models such as BERT, GPT, and RoBERTa. The comparative method is employed to identify the advantages and disadvantages of each approach in the context of specific tasks related to literary analysis. Additionally, the experimental method is applied to evaluate the effectiveness of automated models in identifying the concept of «tolerance» in texts from respected media sources such as the New York Times, BBC News, and The Guardian, in comparison with manual lexical analysis.

Traditional methods of concept extraction in literary texts rely on manual work by researchers and the use of lexical analysis. These approaches are based on in-depth reading of the text, identifying key words, analyzing their contextual meanings, and examining the relationships between them. Lexical analysis, which serves as the foundation of these methods, includes the identification of word frequency, their distribution across the text, and the detection of thematic dominants. Such methods allow for the identification of core semantic units within a text, but they require significant cognitive effort and time [9]. Manual work by the researcher has always been a central part of conceptual analysis. Researchers rely on their knowledge of context, history, culture, and the stylistic features of the work in order to identify concepts. The manual approach allows for the consideration of nuances in the text that are difficult to capture using automated tools [8]. Lexical analysis also involves creating lists of key words or dictionaries, which become the basis for further text analysis [19].

Let us consider the use of graphs to represent knowledge extracted from a text in the form of a semantic network. As input for building the graph, we will use the following paragraph: «Artificial intelligence systems (AIS) are a field of computer science that

studies methods for creating intelligent systems. The main areas of AIS include machine learning, natural language processing, and artificial neural networks. Machine learning uses methods for data analysis, while natural language processing enables machines to understand human language». From this text, the following concepts can be identified: artificial intelligence (AI), computer science, intelligent systems, machine learning, natural language processing, neural networks, methods, data analysis, and human language. Logical connections can be traced between these concepts as follows:

Artificial intelligence systems → are a part of → Computer science;

Artificial intelligence systems → include → Machine learning and Natural language processing;

Machine learning → uses → Methods;

Methods → are applied for → Data analysis;

Natural language processing → works with → Human language.

Based on these concepts and the relationships between them, we can build a directed graph, where the nodes represent the concepts and the edges represent the logical connections.

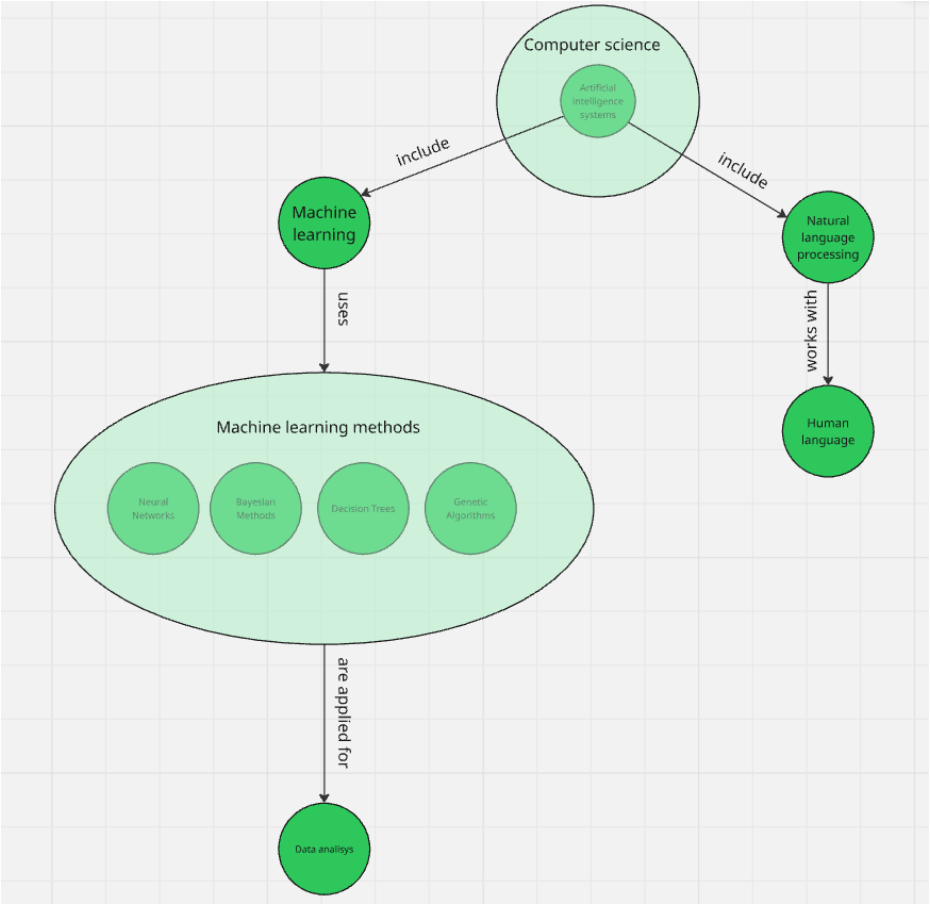

Fig. 1. Knowledge graph based on extracted concepts

This approach makes it possible to visualize knowledge representation in the form of a graph, which simplifies the process of identifying connections between concepts. Semantic graphs are widely used in information retrieval, ontology construction, and text analysis [29].

Automated methods for concept identification leverage large language models such as BERT, GPT, RoBERTa, and T5, which are capable of detecting concepts in complex texts thanks to their transformer architecture. The process of automatic concept detection begins with tokenization – the stage where the text is divided into individual words or phrases (tokens). This can be a simple split based on spaces or more advanced approaches that take into account morphological and syntactic features of the language. In the next stage, the tokens are transformed into vector representations (embeddings). These embeddings carry semantic and syntactic information, allowing the models to better understand the contextual meaning of each token [28]. The next step involves identifying semantic dependencies between words using attention mechanisms, which allow the models to effectively consider context and determine the importance of connections between words, particularly in complex sentences with ambiguous structures or polysemous terms. In the final stage, clustering methods and additional attention layers are used to extract key concepts. Clustering helps group semantically similar tokens into conceptual clusters, enabling clearer separation between different concepts and highlighting the central themes of the text [1].

A prominent example of a transformer architecture used for concept extraction tasks is BERT (Bidirectional Encoder Representations from Transformers), developed in 2018 by a team of researchers at Google led by Jacob Devlin. BERT quickly became a standard in various natural language processing tasks such as text classification, sentiment analysis, automatic concept extraction, and more. Its success contributed to the broad development of transformer-based models and had a major influence on the next generations of large language models, including RoBERTa, ALBERT, and GPT. BERT distributes «attention focus» across important parts of the text to improve understanding, and it is a bidirectional model, which means it takes into account context from both the left and the right. This leads to a much better understanding of contextual connections between words in a text, which is essential for accurately identifying key concepts. The text moves from raw tokens to a final vector representation, which the model uses to determine the core concepts in the text:

1. Text tokenization – the process of breaking down the text into tokens (words or subwords).
2. Embedding generation – transforming tokens into vector representations that preserve semantic and syntactic features.
3. Use of the attention mechanism – allows the model to determine the importance of each token based on the full sentence context.
4. Multi-layer encoding – the model consists of several layers that refine token representations using precomputed weight values.
5. Classification and concept extraction – the final vector representations are used to identify concepts or perform other NLP tasks.

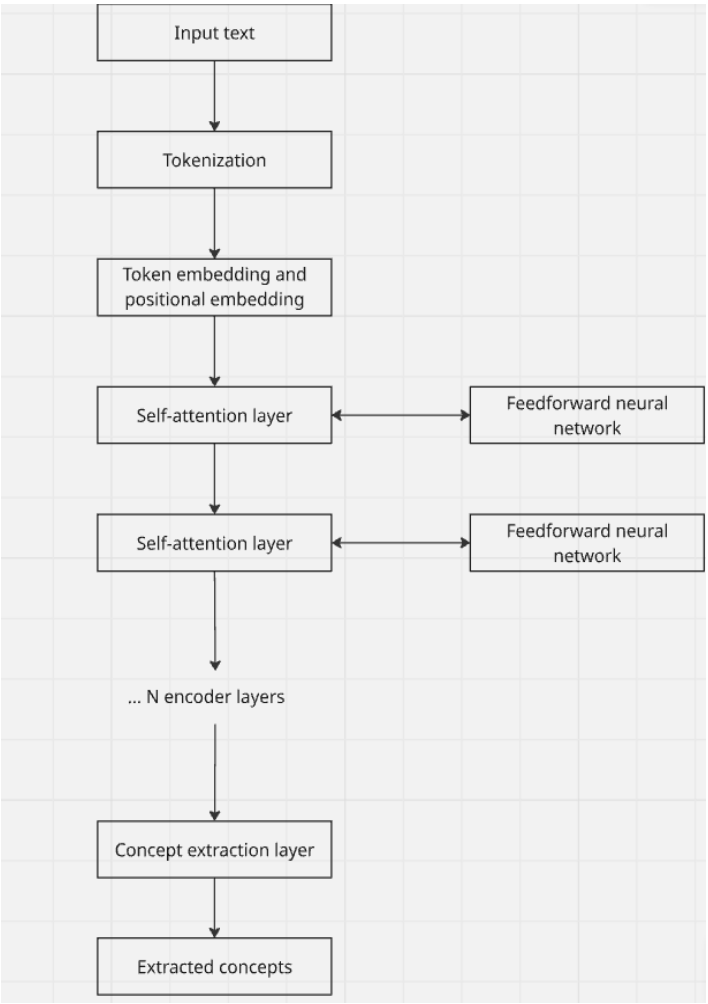

Fig. 2. BERT model workflow

For instance, in analyzing the sentence

«Neural networks are the foundation of deep learning, which is used for pattern recognition, natural language processing, and autonomous systems»,

BERT employs the WordPiece Tokenization algorithm to break down words into subwords or tokens. The self-attention mechanism enables it to consider interdependencies between words and understand their roles within the sentence context. BERT applies the Named Entity Recognition (NER) approach to automatically identify key concepts, recognizing «Neural Networks» and «Deep Learning» as central concepts, with «Pattern Recognition» identified as one of their applications [15].

Concept detection can be further optimized using clustering of vector representations generated by large language models such as BERT [1]. This approach is particularly effective for analyzing large volumes of textual data, as it allows the identification of hidden structures and relationships between various concepts. The clustering process begins with converting the text into vector representations using the BERT model, which generates vectors containing deep semantic information about words and phrases [16]. These vectors are then grouped based on thematic similarity using clustering algorithms such as DBSCAN or k-means.

The automation of concept extraction is also closely linked to ontology-based modeling. Ontologies allow for the structuring of knowledge and simplify the analysis of semantic connections between elements within a text [29]. For example, to formalize knowledge about concepts in the domain of «Artificial Intelligence Systems», it is necessary to identify classes – the core entities of the domain – such as «Artificial Intelligence Systems», «Machine Learning», and «Natural Language Processing». Each class can be assigned attributes that characterize it. For example, for «Artificial Intelligence Systems», attributes may include the domain of application (medicine, automobiles, robotics, finance) and methods used (machine learning, logic, heuristics). The class «Machine Learning» can be described by attributes such as types of learning (reinforcement, supervised, unsupervised) and key methods (neural networks, Bayesian methods, decision trees), while «Natural Language Processing» can be described using attributes like «Data Type» (text, speech) and «Core Tasks» (Syntactic Analysis, Semantic Analysis). Relationships between concepts include «is-a» (e.g., machine learning is a subtype of artificial intelligence systems), «part-of» (e.g., semantic analysis is a part of natural language processing), and «uses» (e.g., natural language processing uses deep learning).

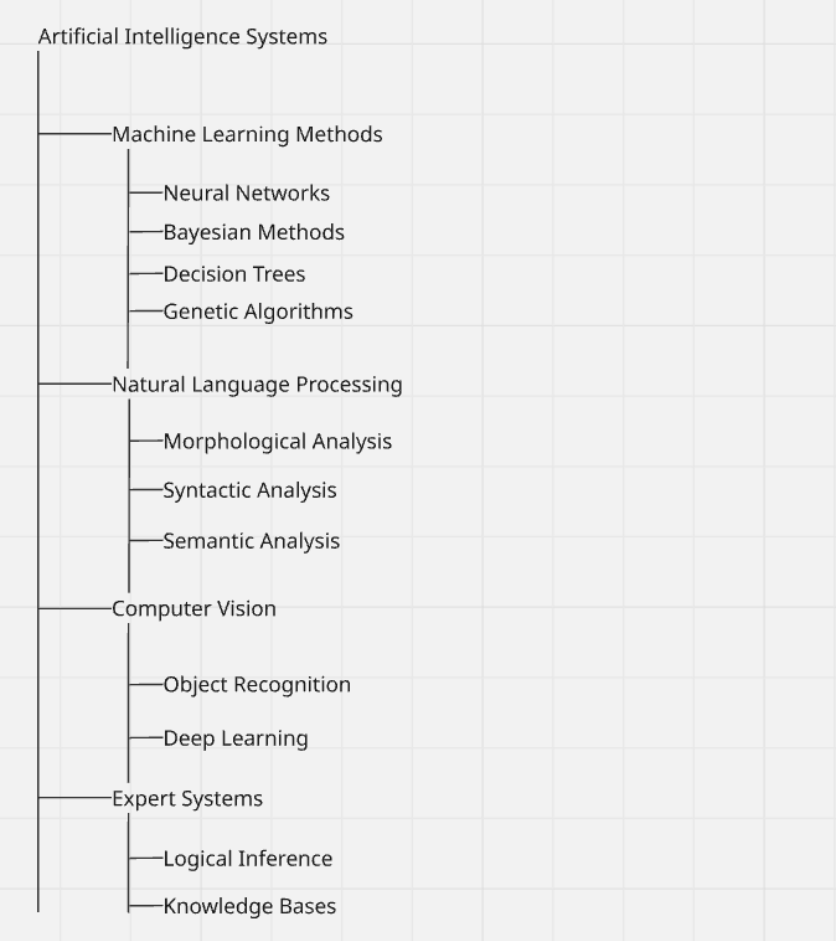

Fig. 3. Tree-like Structure of Ontology Classes

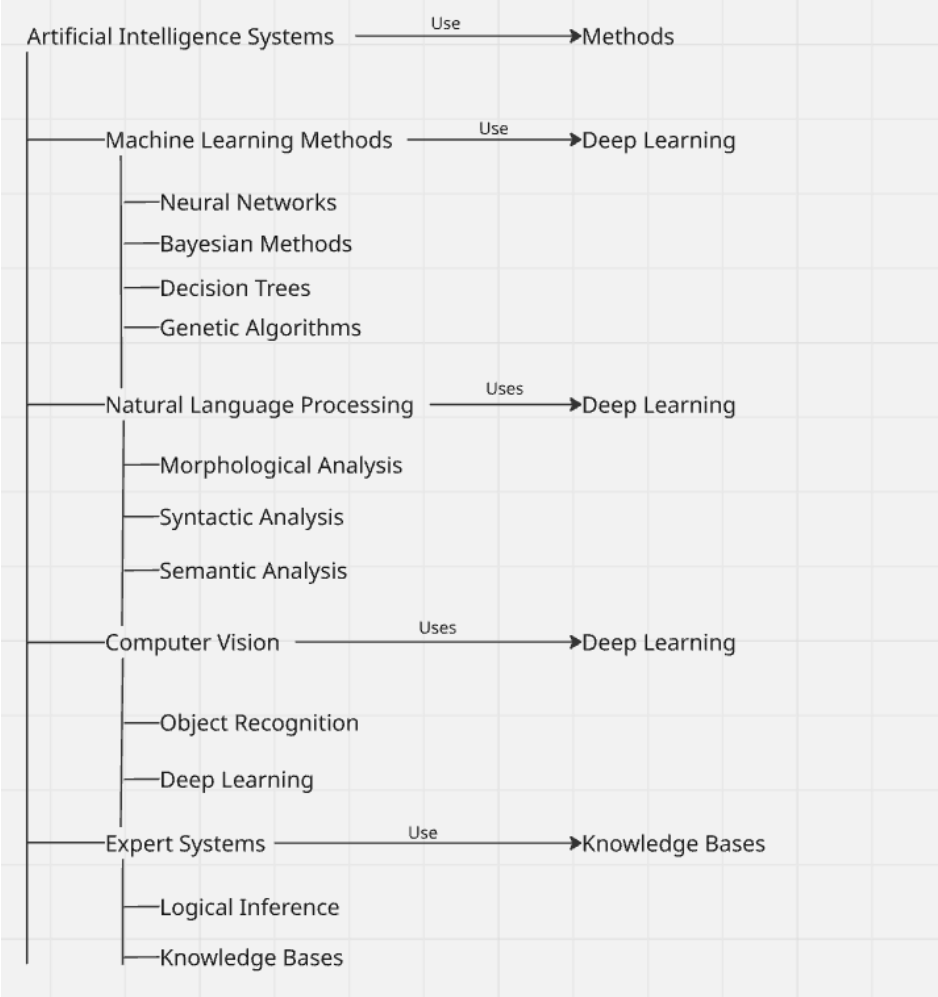

Fig. 4. Ontology of the subject area «Artificial Intelligence Systems»

At the current stage of concept extraction automation, knowledge graphs and combined approaches that integrate the capabilities of large language models and ontologies are widely used. In the study [26], a model is presented that enables the creation of ontologies based on text analysis using advanced concept extraction methods. This approach combines the structured nature of ontologies with the flexibility of large language models, opening new possibilities for the analysis of complex literary phenomena.

Advanced technical aspects include the use of a dual-agent setup for progressively refining results, which significantly improves the quality of extracted concepts [13]. For example, in analyzing a literary text like «The entire land was covered in snow, and the stars twinkled high in the sky with a cold light», the first agent proposes preliminary hypotheses about symbolic meanings (e.g., «snow» as loneliness, «stars» as hope), while the second agent refines these within the broader literary context. Additionally, pretraining models on domain-specific datasets, such as classical literature, enhances their ability to recognize literary concepts and metaphors [15; 26].

The experimental analysis focuses on the concept of «tolerance» in 400 sentences collected from The New York Times, BBC News, and The Guardian. Manual lexical

analysis examines each sentence to identify linguistic means used to express the concept, including synonyms, antonyms, contextual usage, and stylistic features. Automated analysis, using the GPT-4.5 model, performs a similar lexical analysis, identifying concepts, synonyms, antonyms, and stylistic features, and classifying usages based on connotation (positive, negative, neutral) and thematic domains (social issues, politics, culture). The results of both approaches are compared to assess their effectiveness.

The comparative analysis of traditional and automated methods for concept extraction yielded significant findings regarding their application to literary and media texts. In the study of the concept of «tolerance» in contemporary English-language media, manual lexical analysis was conducted on 400 sentences collected from The New York Times, BBC News, and The Guardian. Each sentence was examined in detail to identify the linguistic means used to express the concept of tolerance, including synonyms, antonyms, contextual usage, and stylistic features. The analysis revealed a diversity of ways in which the concept is verbalized, expressed through direct references to tolerance, synonymous expressions, and metaphorical constructions. It was found that, depending on the context, tolerance carries either positive or negative connotations. For example, in the context of social issues, it often appears with a positive tone, whereas in political discourse, it may acquire negative overtones. Stylistic patterns specific to different media outlets were also identified, showing that different publications use distinct stylistic strategies to convey the concept of tolerance, reflecting their editorial policies and target audiences [12].

A parallel analysis using the large language model GPT-4.5 was performed on the same 400 sentences from The New York Times, BBC News, and The Guardian. The model identified specific ways in which the concept of tolerance is verbalized in contemporary media discourse, examining synonyms, antonyms, contextual usage, and stylistic features. The classification of usages was performed based on connotation (positive, negative, neutral) and thematic domains (social issues, politics, culture). For instance, in The Guardian article [17], the concept is clearly marked and discussed in the context of religious beliefs and freedom of speech:

«A key understanding of tolerance is the willingness to accept ideas or practices that we might despise or disagree with but recognise are important to others».

In the article «Amid Mosul's Ruins, Pope Denounces Religious Fanaticism: Live Updates» [2], the concept of tolerance is presented as part of a broader image of interfaith unity during Pope Francis's visit to Iraq:

«It was a day meant to convey images of religious unity and tolerance».

In contrast, a BBC News article [23], the word tolerance appears in a scientific context – desiccation tolerance – unrelated to the social or moral concept:

«Farrant, now a professor of desiccation tolerance at the University of Cape Town, has been studying these unusual plants for over three decades».

The analysis confirmed the contextual dependence of the concept and its capacity to carry both positive and negative connotations, consistent with findings from manual analysis [12].

Further examples from the GPT-4.5 analysis include The Guardian article [7], where tolerance is examined within a religious context:

«Dragged into the politicisation of identity, tolerance has become a form of 'polite etiquette', argues Frank Furedi in a new book».

In The New York Times article [22], the concept is used in a comparative context:

«The tolerance that we find in the armed forces, we don't find it outside," said Second Master Anouar...».

A BBC News article [24] presents tolerance as part of an institutional stance against discrimination:

«We have a zero-tolerance policy towards discriminatory behaviour so as soon as issues were pointed out to us, we acted swiftly to protect those affected and show a strong line against poor behaviours that do not reflect the club's values».

Additionally, The Guardian [13] links tolerance to Voltaire's philosophical treatise:

«A bestseller in the wake of Charlie Hebdo, this 18th-century criticism of religious violence is still relevant today».

The New York Times [21] demonstrates the use of tolerance within the fixed expression "zero tolerance policy," which carries a negative connotation:

«The 'zero tolerance' policy was supposed to serve as a deterrent to families traveling with children».

A BBC News article [5] uses tolerance in a technical context, referring to measurement allowances in motorsports:

«That includes a 0.25mm tolerance permitted on the basis of the short notice involved that will be removed for the following race in Japan on 4–6 April, reducing the permitted gap during the test to 0.5mm».

The automated analysis by GPT-4.5 took only 5 minutes – 4 minutes and 30 seconds for query formulation and 30 seconds for actual generation – compared to the many hours required for manual research. This demonstrates a significant advantage in processing speed for large-scale text analysis. The results of the automated analysis were consistent with the manual analysis [12], confirming the diversity in the verbalization of the concept of tolerance and its context-dependent connotations. However, more recent articles analyzed by GPT-4.5 showed a deeper discussion of tolerance in relation to freedom of speech and religious beliefs, potentially reflecting ongoing societal debates.

In addition to the media analysis, the application of the BERT model for concept detection was tested on a scientific text example: «Neural networks are the foundation of deep learning, which is used for pattern recognition, natural language processing, and autonomous systems». BERT's WordPiece Tokenization algorithm broke down the text into subwords or tokens, and the self-attention mechanism identified interdependencies between words. As a result, BERT recognized «Neural Networks» and «Deep Learning» as central concepts, with «Pattern Recognition» identified as one of their applications. This demonstrates the model's ability to reveal the conceptual structure of analyzed texts effectively.

The comparative analysis of traditional and modern automated methods for concept extraction in literary and media texts reveals distinct strengths and limitations for each approach, providing insights into their applicability and potential for integration. Traditional methods, such as manual lexical analysis and semantic interpretation, offer a deep understanding of the text by leveraging expert knowledge and contextual awareness. These methods are particularly effective for analyzing complex texts rich in symbolism, metaphors, and allusions, where cultural, historical, and stylistic nuances play a critical role [8; 9]. For instance, the manual analysis of the concept of «tolerance» in 400 sentences from The New York Times, BBC News, and The Guardian demonstrated the ability to capture subtle variations in verbalization, such as positive or negative connotations and media-specific stylistic strategies [12]. However, their labor-intensive nature, subjectivity, and limited scalability make them impractical for large text corpora, as they demand extensive human resources and time [29].

In contrast, large language models like GPT-4.5 and BERT provide significant advantages in speed and scalability, enabling rapid processing of large datasets. The automated analysis of the same 400 sentences using GPT-4.5, completed in just 5 minutes compared to hours for manual analysis, underscores the efficiency of these models for large-scale studies. The consistency of GPT-4.5's findings with manual analysis, as evidenced by the identification of diverse verbalizations of «tolerance» (e.g., direct references, metaphors, and context-dependent connotations), highlights the models' ability to analyze semantic relationships and context effectively [13]. Similarly, BERT's successful identification of concepts like «Neural Networks» and «Deep Learning» in a scientific text example demonstrates its capacity to reveal conceptual structures in complex texts. The integration of large language models with ontologies further enhances accuracy by reducing interpretive ambiguity, as ontologies provide structured knowledge and facilitate clearer identification of relationships between concepts [18].

The results of this study align with prior research, such as Hlibovska [12], which also noted the contextual dependence and varied verbalization of «tolerance» in media discourse. The automated analysis not only confirmed Hlibovska's findings but also highlighted emerging societal debates, such as those on freedom of speech and religious beliefs, demonstrating the models' ability to capture evolving trends. Nonetheless, the subjective depth provided by manual analysis remains unmatched for certain tasks, particularly when interpreting culturally specific or author-intended meanings [8].

Despite these strengths, automated methods face notable challenges. One significant issue is the occurrence of «hallucinations», where large language models generate irrelevant or incorrect concepts due to limitations in training data or insufficient domain specialization [20]. For example, while GPT-4.5 accurately identified the concept of «tolerance» in media texts, its performance could be compromised in literary texts with layered metaphors or cultural allusions, where deeper contextual understanding is required. This limitation is particularly relevant for literary analysis, where texts often contain stylistic and symbolic elements that automated models may struggle to

interpret fully. The study suggests that ontology-based approaches can mitigate such risks by providing factual consistency and structural clarity, but further refinement of models is necessary to handle the complexity of literary texts [20].

The practical applications of automatic concept detection methods are diverse and in high demand across many fields. One key application is the automatic construction of ontologies in the development of information systems [6]. This process involves creating structured semantic networks that represent the relationships and hierarchy of concepts within specific knowledge domains. Another important area is the processing of scientific texts, where automatic detection of new concepts and their relationships accelerates analysis of articles and identification of research trends. Additionally, in semantic search, conceptual vector representations enable more accurate search systems that retrieve information based on meaning and context, increasing relevance in fields like medicine and business analytics [16].

The integration of traditional and automated approaches emerges as a promising solution to leverage the strengths of both. Automated methods, such as those using BERT or GPT-4.5, are ideal for initial large-scale analysis, quickly structuring large text corpora and identifying key concepts. Subsequent manual analysis by experts can refine these findings, correcting potential errors (e.g., «hallucinations») and providing nuanced interpretations of complex literary phenomena. For instance, combining BERT's clustering capabilities with expert validation could enhance the accuracy of concept extraction in texts rich in metaphors or allusions. The use of ontologies, as demonstrated in the study's ontological modeling of «Artificial Intelligence Systems» further supports this combined approach by structuring knowledge and reducing ambiguity [29].

The prospects for further research cover several important directions. First, it is essential to continue improving and adapting modern automated analysis methods such as large language models, including BERT, GPT, RoBERTa, and ALBERT. This involves a detailed examination of their capabilities and limitations in specific literary tasks, particularly in analyzing symbolic, metaphorical, and stylistic levels of texts. Of particular interest is the study of combined approaches that integrate traditional methods with modern automated techniques. Future research should develop clear guidelines and algorithms to ensure effective collaboration between human experts and computer systems. Another important direction is the development of specialized ontological models aimed at structuring knowledge within specific literary genres or traditions. Exploring the interaction between large language models and ontologies has the potential to greatly enhance the accuracy and contextual relevance of automated analysis. It is also important to investigate the possibilities of automatically detecting and classifying symbolic and metaphorical constructions in literary texts using modern natural language processing technologies. Further research may help determine criteria and parameters for assessing the accuracy of automated systems and interpreting their outputs, considering the contextual, stylistic, and cultural characteristics of literary works. It also should focus on adapting large language models to literary contexts, particularly

by pretraining them on specialized literary corpora to improve their ability to detect symbolic and metaphorical elements [25]. Additionally, creating literary ontologies that account for cultural, historical, and intertextual factors could significantly enhance the contextual relevance of automated analyses. These advancements would enable a more comprehensive approach to conceptual analysis, balancing the efficiency of automation with the interpretive depth of traditional methods.

Conclusions. The conducted analysis of traditional and modern methods for concept extraction in literary texts allows for several important conclusions to be drawn. First, traditional methods – such as manual lexical analysis and semantic interpretation – are characterized by a high level of interpretive depth and the ability to consider broad cultural, historical, and literary contexts. These methods are indispensable when analyzing individual literary works that are rich in metaphors, symbolism, and allusions. However, their major limitations lie in their labor-intensive nature, dependence on the subjectivity of the researcher, and limited applicability to large text corpora.

On the other hand, modern automated methods – particularly large language models based on transformer architecture – demonstrate significant advantages in terms of scalability, processing speed, and the ability to account for deep semantic and contextual dependencies. Their effectiveness in identifying concepts, even in complex and multi-layered texts, has been confirmed by practical examples. Special attention should be given to the integration of large language models with ontological frameworks, which helps reduce interpretive ambiguity and improve the accuracy of concept extraction. The use of ontologies enables more efficient structuring of knowledge and clearer identification of relationships between concepts, which is especially relevant in literary analysis, where it is essential to consider stylistic, cultural, and historical characteristics of texts.

A practical study involving both manual lexical analysis and automated analysis using a large language model confirmed the diversity of verbalizations of the concept tolerance in publications such as The New York Times, BBC News, and The Guardian. Specifically, it was shown that the concept of tolerance is expressed through a variety of lexical means, including direct mentions, synonymous expressions, and metaphorical constructions. It was also found that, depending on the context, tolerance acquires positive, negative, or neutral connotations. At the same time, the automated approach significantly reduced the time required for analysis, demonstrating the feasibility of efficient large-scale processing that would be nearly impossible to perform manually due to the high demand for resources. Special attention should be given to the integration of large language models with ontological frameworks, which helps reduce interpretive ambiguity and improve the accuracy of concept extraction. The use of ontologies enables more efficient structuring of knowledge and clearer identification of relationships between concepts, which is especially relevant in literary analysis, where it is essential to consider stylistic, cultural, and historical characteristics of texts

Nevertheless, modern methods are not without challenges. One notable issue is the occurrence of so-called «hallucinations», in which large language models generate

inaccurate or irrelevant concepts. This calls for continued research and improvement of the models, as well as the application of additional validation mechanisms, particularly those based on ontological methods.

Therefore, the optimal solution is the integration of traditional and modern automated approaches. It is recommended to use automated methods for the initial large-scale analysis of texts, and traditional approaches for subsequent detailed interpretive analysis of individual concepts and complex literary phenomena. This combined approach makes it possible to obtain the most accurate, in-depth, and reliable results in literary analysis. Further research should focus on improving automated methods, particularly by pretraining large language models on specialized literary corpora and developing literary ontologies that account for cultural, historical, and intertextual factors, to enhance the accuracy and contextual relevance of concept extraction.

BIBLIOGRAPHY

1. Aggarwal C.C., Zhai C. A survey of text clustering algorithms. In: *Mining Text Data*. New York: Springer, 2012. P. 77–128. DOI: https://doi.org/10.1007/978-1-4614-3223-4_4.

2. Amid Mosul's ruins, Pope denounces religious fanaticism: Live updates. *The New York Times*. 2021.

3. Areshey A.M. Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet. *Expert Systems*. 2024. Vol. 41. P. 1–27. DOI: https://doi.org/10.1111/exsy.13701.

4. Bartalesi V., Meghini C. Using an ontology for representing the knowledge on literary texts: The Dante Alighieri case study. *Semantic Web*. 2017. Vol. 8. P. 385–394. DOI: https://doi.org/10.3233/SW-150198.

5. Benson A. F1 teams face tougher tests on flexi-wings at Chinese GP. *BBC News*. 2025.

6. Brewster C. Ontology learning from text: Methods, evaluation and applications. *Computational Linguistics*. 2006. Vol. 34. P. 569–572.

7. Bunting M. The problem with tolerance. *The Guardian*. 2011.

8. Давидюк Ю.Б. Методика концептуального аналізу художнього тексту. *Мова і культура*. 2014. Вип. 37. № 1. С. 289–293. URL: http://nbuv.gov.ua/UJRN/Mik_2014_17_1_51

9. Фісак І.В. Категорія «концепт» у сучасному науковому дискурсі. *Філологічні науки*. 2014. Вип. 17. С. 69–77. URL: http://nbuv.gov.ua/UJRN/Fil_Nauk_2014_17_12

10. Giglou H.B., D'Souza J., Auer S. LLMs4OL: Large language models for ontology learning. In: *The Semantic Web – ISWC*. Springer, 2023. P. 408–427. DOI: https://doi.org/10.1007/978-3-031-47240-4_22.

11. Giglou H.B., D'Souza J., Enge F. LLMs4OM: Matching ontologies with large language models. *ESWC 2024 Special Track on LLMs for Knowledge Engineering*. 2024. P. 23–34. DOI: https://doi.org/10.13140/RG.2.2.10832.42240.

12. Глібовська А.А. Відтворення національно-культурних особливостей вербалізації концепту «толерантність» у сучасній мові ЗМІ під час перекладу з англійської мови на українську. Київ, 2020.

13. Hu Y., Liu D., Wang Q. Automating knowledge discovery from scientific literature via LLMs: A dual-agent approach with progressive ontology prompting. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2409.00054.

14. Lezard N. Treatise on Tolerance by Voltaire review – An attack on fanaticism. *The Guardian*. 2016.

15. Liu Y., Ott M., Nman G. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1907.11692.

16. Mahboub A., Zater M. E., Al-Rfooh B. Evaluation of semantic search and its role in retrieved-augmented-generation (RAG) for Arabic language. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.18350.

17. Malik K. Ideas can be tolerated without being respected. The distinction is key. *The Guardian*. 2020.

18. Mukanova A., Milosz M., Dauletkaliyeva A. LLM-powered natural language text processing for ontology enrichment. *Applied Sciences*. 2024. Vol. 14. P. 5860–5875. DOI: https://doi.org/10.3390/app14135860.

19. Михайлюк А., Михайлюк О., Пилипчук О. Формування лінгвістичної онтології на базі структурованого енциклопедичного ресурсу. *Радіоелектронні і комп'ютерні системи*. 2012. Вип. 4. С. 81–89. URL: http://nbuv.gov.ua/UJRN/recs_2012_4_14

20. Nananukula N., Kejriwala M. HALO: An ontology for representing and categorizing hallucinations in large language models. In: *SPIE Defense + Commercial Sensing (DCS 2024)*. 2024. P. 1–15. DOI: https://doi.org/10.1117/12.3014048.

21. Nixon R. «Zero tolerance» immigration policy surprised agencies, report finds. *The New York Times*. 2018.

22. Onishi N., Meheut C. In France's military, Muslims find a tolerance that is elusive elsewhere. *The New York Times*. 2021.

23. Riley A. Resurrection plants: The drought-resistant «zombie plants» that come back from the dead. *BBC News*. 2025.

24. Sacked Bradburn fined for discriminatory comments. *BBC News*. 2025.

25. To H.Q., Liu M. Towards efficient large language models for scientific text: A review. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2408.10729.

26. Toro S., Anagnostopoulos A.V., Bello S.M. Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). *Journal of Biomedical Semantics*. 2024. Vol. 15, No. 19. DOI: https://doi.org/10.1186/s13326-024-00317-z.

27. Vaswani A., Shazeer N., Parmar N. Attention is all you need. *arXiv*. 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762.

28. Zhenzhong L., Chen M., Goodman S. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1909.11942.

29. Zulkipli Z.Z., Maskat R., Teo N. H.I. A systematic literature review of automatic ontology construction. *Indonesian Journal of Electrical Engineering and Computer Science*. 2022. Vol. 28. P. 878–889. DOI: https://doi.org/10.11591/ijeecs.v28.i2.pp878-889.

REFERENCES

1. Aggarwal, C.C., & Zhai, C. (2012). *A survey of text clustering algorithms*. https://doi.org/10.1007/978-1-4614-3223-4_4

2. Amid Mosul's ruins, Pope denounces religious fanaticism: Live updates. (2021, March 7). *The New York Times*.

3. Areshey, A.M. (2024). Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet. *Expert Systems, 41*, 1–27. https://doi.org/10.1111/exsy.13701

4. Bartalesi, V., & Meghini, C. (2017). Using an ontology for representing the knowledge on literary texts: The Dante Alighieri case study. *Semantic Web, 8*, 385–394. https://doi.org/10.3233/SW-150198

5. Benson, A. (2025, March 18). F1 teams face tougher tests on flexi-wings at Chinese GP. *BBC News*.

6. Brewster, C. (2006). Ontology learning from text: Methods, evaluation and applications. *Computational Linguistics, 34*, 569–572.

7. Bunting, M. (2011, September 5). The problem with tolerance. *The Guardian*.

8. Davydiuk, Y.B. (2014). Metodyka kontseptualnoho analizu khudozhnioho tekstu (Methodology of conceptual analysis of literary texts). *Mova i kul'tura*, 37, 289–293. Retrieved from http://nbuv.gov.ua/UJRN/Mik_2014_17_1_51

9. Fisak, I.V. (2014). Katehoriia «kontsept» u suchasnomu naukovomu dyskursi (The category of 'concept' in contemporary scientific discourse). *Filolohichni nauky*, 17, 69–77. Retrieved from http://nbuv.gov.ua/UJRN/Fil_Nauk_2014_17_12

10. Giglou, H.B., D'Souza, J., & Auer, S. (2023). LLMs4OL: Large language models for ontology learning. In *The Semantic Web – ISWC* (pp. 408–427). https://doi.org/10.1007/978-3-031-47240-4_22

11. Giglou, H.B., D'Souza, J., & Enge, F. (2024). LLMs4OM: Matching ontologies with large language models. *ESWC 2024 Special Track on LLMs for Knowledge Engineering*, 23–34. https://doi.org/10.13140/RG.2.2.10832.42240

12. Hlibovska, A.A. (2020). *Vidtvorennia natsionalno-kulturnykh osoblyvostei verbalizatsii kontseptu «tolerantnist» u suchasnii movi ZMI pry perekladi z anhliiskoi movy na ukrainsku* (Reproduction of national and cultural features of the verbalization of the concept of «tolerance» in modern media language when translating from English into Ukrainian). Kyiv.

13. Hu, Y., Liu, D., & Wang, Q. (2024). Automating knowledge discovery from scientific literature via LLMs: A dual-agent approach with progressive ontology prompting. *arXiv*. https://doi.org/10.48550/arXiv.2409.00054

14. Lezard, N. (2016, October 4). *Treatise on Tolerance by Voltaire review – An attack on fanaticism*. *The Guardian*.

15. Liu, Y., Ott, M., & Nman, G. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

16. Mahboub, A., Zater, M.E., & Al-Rfooh, B. (2024). Evaluation of semantic search and its role in retrieved-augmented-generation (RAG) for Arabic language. *arXiv*. https://doi.org/10.48550/arXiv.2403.18350

17. Malik, K. (2020, December 13). Ideas can be tolerated without being respected. The distinction is key. *The Guardian*.

18. Mukanova, A., Milosz, M., & Dauletkaliyeva, A. (2024). LLM-powered natural language text processing for ontology enrichment. *Applied Sciences, 14*, 5860–5875. https://doi.org/10.3390/app14135860

19. Mykhailiuk, A., Mykhailiuk, O., & Pylypchuk, O. (2012). Formuvannia lingvistychnoii ontolohii na bazistrukturovanoho entsyklopedychnoho resursu (Formation of a linguistic ontology based on a structured encyclopedic resource). *Radioelektronni i komp'iuterni systemy*, 4, 81–89. Retrieved from http://nbuv.gov.ua/UJRN/recs_2012_4_14

20. Nananukula, N., & Kejriwala, M. (2024). HALO: An ontology for representing and categorizing hallucinations in large language models. In *SPIE Defense + Commercial Sensing (DCS 2024)* (pp. 1–15). https://doi.org/10.1117/12.3014048

21. Nixon, R. (2018, October 24). 'Zero tolerance' immigration policy surprised agencies, report finds. *The New York Times*.

22. Onishi, N., & Meheut, C. (2021, June 26). In France's military, Muslims find a tolerance that is elusive elsewhere. *The New York Times*.

23. Riley, A. (2025, March 19). Resurrection plants: The drought-resistant «zombie plants» that come back from the dead. *BBC News*.

24. Sacked Bradburn fined for discriminatory comments. (2025, March 20). *BBC News*.

25. To, H. Q., & Liu, M. (2024). Towards efficient large language models for scientific text: A review. *arXiv*. https://doi.org/10.48550/arXiv.2408.10729

26. Toro, S., Anagnostopoulos, A.V., & Bello, S.M. (2024). Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI). *Journal of Biomedical Semantics, 15*(19). Retrieved from https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-024-00317-z

27. Vaswani, A., Shazeer, N., & Parmar, N. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

28. Zhenzhong, L., Chen, M., & Goodman, S. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*. https://doi.org/10.48550/arXiv.1909.11942

29. Zulkipli, Z.Z., Maskat, R., & Teo, N.H.I. (2022). A systematic literature review of automatic ontology construction. *Indonesian Journal of Electrical Engineering and Computer Science, 28*, 878–889. https://doi.org/10.11591/ijeecs.v28.i2.pp878-889

Стаття надійшла: 05.07.2025
Прийнято: 08.08.2025
Опубліковано: 02.12.2025