DOI: https://doi.org/10.15276/hait.08.2025.9
UDC 004.91

From classification to taxonomy: Automated structuring of vehicle repair names in multilingual corpora

Sergii V. Mashtalir[1)]
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
Oleksandr V. Nikolenko[2)]
ORCID: https://orcid.org/0000-0002-6422-7824; oleksandr.nikolenko@uzhnu.edu.ua. Scopus Author ID: 59739709200
[1)] Kharkiv National University of Radio Electronics, 14, Nauky Ave. Kharkiv, 61166, Ukraine
[2)] Uzhhorod National University, 14, University Str. Uzhhorod, 88000, Ukraine

ABSTRACT

This study introduces and rigorously validates a hybrid, five-stage Natural Language Processing pipeline that transforms unstructured, bilingual repair-order text into fully navigable, hierarchical action taxonomy – bridging the gap between flat keyword classification and business-grade knowledge organization. Addressing the limitations of both traditional and modern Natural Language Processing methods in technical, noisy, and domain-specific datasets, the proposed methodology integrates advanced lemmatization, manual core dictionary creation, semantic filtering, transformer-based classification, and embedding-driven clustering. Building on advanced Ukrainian lemmatization, dynamic semantic filtering, multilingual sentence embeddings, and density clustering, the pipeline systematically overcomes the noise, code-switching, and "long-tail" rarity that typify real-world automotive datasets. Tested on a corpus of over 4.3 million service records, the approach achieves over 92 % cluster coherence with minimal manual annotation. The resulting taxonomy unlocks four immediate industrial benefits: enterprise-wide repair analytics and benchmarking across branches and brands; intent-aware chatbots capable of precise service triage and automated quotation; inventory and workforce optimization through fine-grained job statistics; and a practical blueprint for industry-level standardization of repair nomenclature and data exchange. In sum, the work demonstrates that combining minimal expert input with modern embedding techniques and density clustering can automate taxonomy induction at industrial scale, setting a new benchmark for digital transformation initiatives that depend on accurate structuring of noisy technical language.

Keywords: Natural Language Processing; taxonomy induction; semantic clustering; machine learning; data analysis; applied intelligent systems; data-driven automation; knowledge organization; business process automation

***For citation*: Mashtalir S. V., Nikolenko O. V. "From classification to taxonomy: Automated structuring of vehicle repair names in multilingual corpora". *Herald of Advanced Information Technology*. 2025; Vol. 8 No. 2: 151–163. DOI: https://doi.org/10.15276/hait.08.2025.9**

INTRODUCTION AND PROBLEM STATEMENT

The transition from simple classification to the automated induction of taxonomies and robust clustering is a central challenge in natural language processing, particularly when working with unstructured, noisy, and multilingual technical text [1].

Taxonomy construction – organizing terms and actions into structured hierarchies – underpins knowledge management, semantic search, and process automation across many industries [2], [3], [4], [5], [6]. However, most real-world corpora lack explicit structure, contain vast lexical and syntactic variation, and exhibit the "long tail" phenomenon [7], [8], where many important concepts are rare or novel.

A classic pipeline begins with the classification of individual terms or phrases – determining, for example, whether a string denotes an "action" or a different entity [9]. However, practical impact is only realized when these classified units are then clustered and organized into taxonomies: systems of categories and relationships that enable analytics, automation, and decision-making. Achieving this transition, especially in multilingual, domain-specific corpora, remains a major unsolved problem in applied NLP and information science [10], [11].

While recent advances in deep learning and pretrained language models – such as BERT, LaBSE, multilingual E5, and large-scale transformers – have transformed natural language processing, these models also exhibit critical limitations in highly specialized, technical domains [12], [13].

Despite significant progress in information retrieval and semantic modeling, classical methods such as TF-IDF, bag-of-words, and rule-based dictionaries fall short in these tasks [14]. They are unable to capture deeper semantic relationships, resolve synonymy and polysemy, or handle the morphological and spelling variation inherent to

© Mashtalir S., Nikolenko O., 2025

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0 /deed.uk)

"live" business data [15], [16]. For example, attempts to group actions using TF-IDF typically fragment semantically identical items and fail to identify important but rare "long tail" actions.

General-purpose language models are typically trained on vast, but mostly general, corpora (e.g., Wikipedia, Common Crawl, social media) and thus may lack sufficient exposure to domain-specific terminology, rare abbreviations, and professional jargon encountered in automotive repair data [17], [18]. As a result, they may generate embeddings that do not accurately reflect the semantic relationships between technical action names, leading to suboptimal clustering and misclassification of key concepts. Moreover, even strong multilingual models can struggle with spelling errors, language-switching, and regionally specific slang that is common in real-world service corpora. Fine-tuning these models for technical classification and clustering often requires significant annotated data and expert intervention, which may not always be feasible at scale.

These challenges motivate the development of hybrid pipelines that combine the strengths of modern language models with domain-specific normalization, rule-based pre- and post-processing, and iterative manual refinement – thus enabling the accurate and robust induction of taxonomies in specialized fields like automotive repair [19].

These challenges are acutely present in the automotive service industry, which serves as a paradigmatic case for this research. Despite the rapid growth of digital platforms, there is currently no global standard for the classification or taxonomy of automotive repair actions [20]. This has led to a landscape where even large manufacturers, insurers, and service providers are unable to generate reliable, comparable statistics or automate business processes at scale [21].

The consequences of this fragmentation are substantial for all stakeholders [22]:

– *Vehicle manufacturers* cannot aggregate failure data or benchmark reliability at the level of systems or components.

– *Insurance companies* face ambiguity in repair cost assessment, risk calculation, and fraud detection due to inconsistent action labeling.

– *Repair garages and service providers* lack efficient workflow management, inventory optimization, and price transparency.

– *Car owners and fleet operators* are unable to make informed decisions about current and future maintenance costs, or to benefit from aggregated benchmarking.

Furthermore, with the rise of business automation, conversational AI, and digital self-service (such as chatbots and voice assistants), the ability to accurately detect, extract, and structure repair action names is fundamental for tasks like automated repair estimation, appointment scheduling, and customer support.

For instance, the Ukrainian phrase "заміна масла" ("oil replacement") may appear in a multitude of forms: "замінити масло", "поміняти мастило", "заміна оливи двигуна", etc. Normalization and taxonomy induction must be robust to such diversity, including spelling errors, mixed languages, and professional jargon.

These real-world requirements underscore the need for modern, hybrid pipelines [23] – integrating rule-based, morphological, machine learning, and semantic clustering approaches – that enable the automated transition from unstructured repair data to structured, actionable taxonomies. Such pipelines must deliver not only accurate classification of actions, but also effective clustering and taxonomy induction, supporting analytics, business intelligence, and smart digital services across the automotive sector.

In this work, we present a hybrid approach for the automated classification and taxonomy construction of repair actions in large, primarily bilingual (Ukrainian-Russian) service corpora. This approach is explicitly designed to bridge the gap between unstructured technical text and structured domain knowledge, laying the foundation for a new generation of data-driven analytics, automation, and digital transformation in automotive aftersales.

RESEARCH OBJECTIVE

The primary goal of this research is to design, implement, and evaluate a hybrid Natural Language Processing (NLP) pipeline for automated taxonomy induction of automotive repair actions.

To achieve this goal, the following specific tasks were formulated:

– analyze and preprocess automotive repair datasets, addressing bilingual linguistic noise and diversity;

– develop and validate a morphological and semantic-based extraction approach for identifying repair actions;

– train and evaluate transformer-based semantic classification models, specifically addressing the multilingual and domain-specific nature of the data;

– implement embedding-based semantic clustering methods to group similar repair actions;

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

– construct, validate, and practically apply a structured hierarchical taxonomy of automotive repair actions.

The approach aims to bridge the gap between unstructured, noisy, and linguistically diverse repair records and the structured, hierarchical knowledge required for effective analytics, business automation, and decision support. By moving "from classification to taxonomy," the pipeline is intended to support continuous adaptation to new terminology, robust semantic normalization across languages and domains, and the scalable construction of a living, business-ready taxonomy of repair actions.

CORPUS DESCRIPTION AND REAL-WORLD DATA CHALLENGES

Data source basic characteristics

The foundation of this research is an extensive, real-world corpus reflecting five years of operations in a modern automotive service management system in Ukraine. The experimental dataset comprises 4,391,597 records of repair work names, aggregated from over 1.5 million repair orders collected from 500 garages for 5 years. After duplicate removal, the number of unique rows stands at 652,929, which means that more than 85 % of entries are exact copies – a finding that highlights the high degree of duplication typical for this industry. This redundancy significantly reduces the scale for subsequent processing and lowers computational requirements for classification and taxonomy induction, making such tasks feasible on standard desktop computers.

Typical characteristics of the text data are as follows:

– average length (characters): 32.22;

– average length (words): 3.90.

These figures confirm that the majority of records are highly compact, consisting of short, functional phrases such as "заміна амортизатора" ("shock absorber replacement") or "ремонт супорта" ("caliper repair").

This brevity creates unique challenges for automatic normalization and semantic grouping, reinforcing the importance of an explicit, hierarchical taxonomy.

Multilingual and Noisy Structure

Due to the bilingual documentation practices of Ukrainian service stations, a hybrid language identification procedure was implemented for every entry. This involved both heuristic character-based checks (for distinctive Ukrainian or Russian letters) and the use of the langdetect model for ambiguous cases. The resulting language distribution is shown in *Table 1*.

Ukrainian clearly dominates (>80 % of the corpus). Russian accounts for about 14 %. Small shares classified as Macedonian or Bulgarian are artifacts of automatic language detection on very short or ambiguous texts, reflecting structural similarities across Slavic languages. The share of English or other Western European languages is negligible.

Table 1. **Language distribution of repair records**

Language	Row Count	Share (%)
Ukrainian	3,515,323	80.05
Russian	606,578	13.81
Macedonian	142,579	3.25
Bulgarian	120,599	2.75
Other	6,518	<0.15

Source: **compiled by the authors**

Thus, the corpus presents a robust bilingual (Ukrainian/Russian) structure with isolated instances of other languages, which do not significantly impact the overall quality of analysis.

Text Preprocessing & Normalization

In our previous research [24], we investigated in detail the linguistic obstacles posed by Ukrainian, Russian, and their hybrid vernacular "surzhyk" in technical corpora.

The present work recalls the salient preprocessing stages:

1) language detection employing language-specific characters and langdetect model;

2) case folding to render all text lowercase;

3) normalization, including removal of extraneous symbols and elimination of stop-words;

4) machine translation of every record into ukrainian – the study's reference language – while automatically constructing a russian–ukrainian term dictionary;

5) grammatical error correction driven by the induced dictionary and token statistics;

6) prefix segmentation for domain-specific morphemes (e.g., auto-, electro-, pneumatic-);

7) abbreviation expansion and synonym standardization;

8) lemmatization leveraging comprehensive on-line ukrainian lexical resources;

9) token filtering and truncation to satisfy transformer input-length constraints.

Taxonomy as Both a Necessity and Solution

The linguistic and structural diversity of the dataset – short, ambiguous, and inconsistent entries

– makes taxonomy induction not just a goal, but a necessity. Without a systematic way to group and organize these records, any attempt at analytics, automation, or even search quickly breaks down.

To support model training and evaluation, a reference hierarchical taxonomy was constructed manually, using the "Chassis" category as a pilot. This "gold standard" lexicon includes 350 unique entries and implements a 6-level structure of categorization as shown in Table 2.

Table 2. **Six level structure of the automotive works directory**

Level	Level name	Example Uk	Example En
1	category	шасі	chassis
2	system	гальмівна	brake
3	subsystem	дискова	disc
4	unit	супорт	calipers
5	component	направляюча	guide pin
6	action	змащення	lubrication

Source: **compiled by the authors**

This taxonomy is structured hierarchically, reflecting a logical breakdown of automotive components and related repair actions:

– level 1 (highest): broad vehicle categories (e.g., chassis, engine, electrical system);

– level 2: specific systems within each category (e.g., brake system, suspension, steering);

– level 3: subsystems within each system (e.g., disc brakes, drum brakes, hydraulics);

– level 4: functional units (e.g., caliper, disc, pads);

– level 5: individual components of units (e.g., guide pins, coils, seals);

– level 6 (lowest): specific repair actions (e.g., lubrication, adjustment, replacement).

Each level groups semantically related concepts, ensuring clear hierarchical separation. Synonymous examples, such as "replacement" ("заміна") and "installation" ("установка"), clearly illustrate semantic proximity at the lowest action level. However, within the scope of the current research, we focus exclusively on the "action" level (level 6) of this hierarchical structure. This decision allows us to deeply investigate the semantic nuances, clustering strategies, and multilingual classification of repair actions. Classifications and automated structuring related to higher levels (1 to 5) will be described and analyzed in detail in separate future works.

Focusing exclusively on the 6th level ("action") of the taxonomy is both practical and methodologically sound. Actions (*заміна* – replacement, *ремонт* – repair, *діагностика* – diagnostics) represent the most frequent, linguistically diverse, and semantically complex elements of automotive repair records. Structuring this critical level first establishes a robust semantic foundation necessary for effective automation, analytics, and knowledge management. Properly normalized and clustered actions will significantly simplify future classification efforts at other taxonomy levels, enhancing overall system coherence and usability.

In summary, the size, duplication, linguistic complexity, and compactness of real repair corpora demand automated, data-driven taxonomy induction. The taxonomy is not simply a tool for grouping records – it is the essential infrastructure for scalable analytics, process automation, and business intelligence in the automotive service domain.

The following sections detail how the hybrid pipeline leverages these data realities – moving from the classification of repair actions to the automated structuring and continuous extension of a living, business-ready taxonomy.

HYBRID PIPELINE FOR ACTION EXTRACTION DESIGN AND PREPROCESSING

In line with our research objective – to transition from flat classification to an automated, hierarchical taxonomy of automotive repair actions – we developed a hybrid, multi-step pipeline. This pipeline integrates manual annotation, rule-based linguistic preprocessing, morphological normalization, followed by machine learning-based semantic classification, clustering and taxonomy induction described in later chapters. A step-by-step representation of this pipeline within the broader data-science lifecycle is provided in Fig. 1.

Initial Manual Annotation and Seed Dictionary

We began by manually selecting a small set of approximately 50 common actions, such as "заміна" ("replacement"), "ремонт" ("repair"), "діагностика" ("diagnostics"), and "регулювання" ("adjustment").

These manually annotated actions served as seed terms, providing an initial semantic core to guide the pipeline.

Morphological Action Definition

To expand this core and reliably extract actions, we introduced a morphological definition of an action, comprising two criteria.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

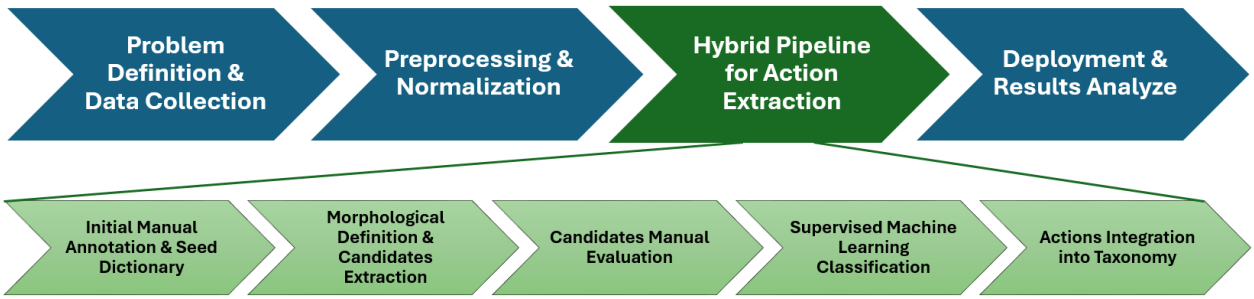


Fig. 1. **Hybrid pipeline for actions extraction within the data-science lifecycle**
Source: **compiled by the authors**

1) *Noun in the nominative singular*

This requirement is addressed through lemmatization (the process of reducing a word to its dictionary form). However, neither standard Python library – Stanza nor lang-uk – provided sufficiently accurate lemmatization for technical Ukrainian texts.

The solution was to leverage the Vesum (ВЕСУМ) – comprehensive electronic dictionary of Ukrainian [25], which covers almost 7 million word forms, thus greatly improving lemmatization accuracy for specialized language directories.

2) *Origin from a verb (verbal noun)*

This requirement posed a greater challenge, as no available resource or dictionary provided etymological data for Ukrainian words at scale.

To approximate this, we relied on the seed dictionary. An embedding vector was computed for each lemma in the corpus. Then, the cosine similarity [26] between the lemma vector and seed dictionary term was calculated. If the cosine similarity with the closest core action exceeded a threshold (0.85), the lemma was considered to be an action.

Candidate Action Extraction through Linguistic Preprocessing and Semantic Filtering

The search for actions in the corpus followed this sequence for each unique record:

– all lemmas in the record were checked for a match against the manual action dictionary;

– if no match was found, the first three lemmas of the record were selected;

– this is motivated by the observation that action lemmas most frequently appear among the first words of repair descriptions (e.g., “Заміна оливи” – “Oil replacement”; “Програмування ECU (Electronic Control Unit)” – “ECU programming”; “Комп’ютерна діагностика” – “Computer diagnostics”);

– each of these three lemmas was checked for correspondence to the action definition (i.e., nominative noun with high semantic similarity to a core action). This filtering was critical, as the corpus contained numerous nouns unrelated to actions but frequently co-occurring in repair descriptions;

– only those lemmas with frequency exceeding 10 occurrences in the corpus were included in the resulting action dictionary, to reduce noise and enhance practical relevance.

Manual Validation and Training Data Expansion

The above process produced 562 new candidate actions. These candidates were then manually reviewed and classified into correct and incorrect action categories. The accuracy at this stage was moderate – only 432 (77 %) of candidates were judged correct.

However, these validated actions expanded the annotated dataset, providing crucial labeled data for training and embedding-based classification model. This model in turn enabled the discovery of actions anywhere in a record, including the so-called “long tail” of rare or novel action terms – discussed in detail in the following chapter.

ACTION CLASSIFICATION AND LONG-TAIL ANALYSIS

The central element of our taxonomy construction pipeline is the accurate semantic classification of candidate terms at the “action” level. Effective classification provides the semantic core essential for subsequent clustering and hierarchical structuring, thereby enabling a smooth transition from raw linguistic data to a well-defined taxonomy.

Machine Learning Classification with Embeddings

To accurately identify and filter genuine action terms, we implemented a supervised machine learning classifier utilizing modern transformer-based sentence embeddings (multilingual-e5-large) [27]. Embeddings were specifically chosen over other deep learning architectures such as RNN and CNN for their ability to capture long-range semantic

dependencies across multilingual data, enabling accurate discrimination of action-related terms from irrelevant nouns and noise [28].

Specifically, transformers demonstrated critical advantages including:

- effective handling of multilingual semantics through pretrained multilingual embeddings;
- robustness to variations and noise in short technical phrases;
- improved semantic clustering performance compared to traditional CNN or RNN approaches.

The multilingual-e5-large embedding model used here follows the original encoder-only transformer architecture (similar to BERT). Key hyperparameters included: 24 transformer encoder layers, embedding dimension of 1024, learning rate of 2e-5 using Adam optimizer [29].

As mentioned in the previous chapter the classifier was initially trained on manually labeled data comprising 562 carefully annotated candidate terms as shown in Table 3.

Table 3. **Extract from manually labeled data**

Action (Uk)	Action (En)	Label
заміна	replacement	action
діагностика	diagnostics	action
ремонт	repair	action
мастило	lubricant	non-action
супорт	caliper	non-action
комплекс	complex (service pack)	non-action

Source: **compiled by the authors**

For the classification of candidate actions in the high-dimensional embedding space, we selected logistic regression due to its suitability for binary classification tasks, particularly when working with semantic embeddings generated by transformer models. These embeddings inherently encode rich semantic information and often render the two classes – "actions" and "non-actions" – linearly or nearly linearly separable, allowing logistic regression to provide an interpretable, robust, and computationally efficient solution that minimizes the risk of overfitting. The semantic compression achieved by the embeddings meant that the classifier could reliably learn a high-quality decision boundary with fewer labeled instances, as evidenced by the model's high accuracy, precision, and recall. Intuitively, this process can be understood as learning a hyperplane in the embedding space that separates action vectors from non-action vectors, leveraging the geometric properties of the space where semantically similar items are clustered together.

To rigorously evaluate the performance of our classifier, we utilized several standard metrics widely adopted in machine learning and information retrieval: accuracy, precision, recall, and F1-score [30].

Accuracy measures the overall proportion of correct predictions, indicating how frequently the classifier made the right decision for both action and non-action terms and is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where *TP* (true positives) and *TN* (true negatives) are the correctly classified examples, while *FP* (false positives) and *FN* (false negatives) represent errors.

Precision for the "action" class quantifies the proportion of predicted actions that were indeed correct, reflecting the model's ability to avoid false positives.

$$Precision = \frac{TP}{TP + FP}$$

Recall for the "action" class represents the proportion of all true actions that the classifier successfully identified, highlighting its sensitivity to genuine actions.

$$Recall = \frac{TP}{TP+FN}.$$

The F1-score provides a harmonic mean of precision and recall, offering a balanced measure that accounts for both false positives and false negatives.

$$F1 - score = \frac{2}{Precision^{-1} + Recall^{-1}} = \\ = \frac{2TP}{2TP+FP+FN}.$$

High values across all these metrics confirm not only the classifier's correctness but also its reliability in detecting rare and ambiguous cases – crucial for robust taxonomy induction in diverse, noisy corpora.

The trained classifier achieved strong performance metrics, demonstrating practical utility in the real-world scenario:

– Accuracy – correct predictions rate: **88.6** %;

– Precision (action) – correct positive predictions rate: **97.4** %;

– Recall (action) – found positives of all actual: **87.4** %;

– F1-score (action) – harmonic mean of precision and recall: **92.2** %.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Handling the "Long-Tail" Phenomenon

A critical challenge in action classification is the "long-tail" distribution, where many important yet infrequent or emerging repair terms occur. In our corpus, thousands of rare actions appeared fewer than 10 times each – terms such as "перезаливання" ("refilling"), "промащування" ("greasing"), and "розкручування" ("unscrewing"). These terms, despite their rarity, can represent essential new or highly specific repair procedures and therefore must not be neglected in taxonomy construction.

Our classifier effectively addressed this issue by leveraging semantic embeddings to classify previously unseen or rarely observed terms accurately. A detailed evaluation demonstrated robust long-tail handling capabilities:

Among 1,816 rare candidate actions (each occurring fewer than 10 times), automatic classification correctly identified more than 90 % of genuine action terms. Conversely, the classifier reliably excluded irrelevant terms mistakenly captured as actions due to linguistic ambiguity or structural similarity.

Integration of Classified Actions into Taxonomy

Classification results directly fed into the subsequent clustering and taxonomy induction stage. Correctly classified actions were semantically grouped into coherent clusters, significantly reducing semantic redundancy and ensuring precise mapping onto higher-level categories. Misclassifications and ambiguous terms were systematically reviewed, minimizing downstream noise.

Thus, robust long-tail classification facilitated the continuous extension and updating of the taxonomy, enabling it to dynamically capture emergent repair actions and adapt to the evolving automotive service landscape.

In summary, the embedding-based classification approach provided a high-accuracy semantic foundation for action taxonomy induction, particularly addressing the critical long-tail problem inherent in technical domains. This robust classification significantly streamlined subsequent taxonomy construction, underpinning the development of a comprehensive, hierarchical, and dynamic knowledge structure necessary for practical automotive analytics and automation.

Having validated and semantically filtered the action lexicon, we proceeded to group actions into semantic clusters and induce a hierarchical taxonomy, as discussed in the next chapter.

CLUSTERING AND TAXONOMY INDUCTION OF ACTION LEXICON

Building on the robust classification described in the previous chapter, we proceeded with clustering actions into semantically coherent groups, enabling structured taxonomy induction.

Three clustering approaches were clearly defined and sequentially applied:

1) Semantic clustering (embedding-based): Hierarchical density-based clustering (HDBSCAN) leveraging multilingual sentence embeddings;

2) Rule-based adjustments: Explicit linguistic rules to merge or split clusters based on specific semantic criteria;

3) Manual expert adjustments: Domain expert review for semantic validation and minor refinements.

Semantic Clustering (Embedding-based)

Using multilingual sentence embeddings (multilingual-e5-large), we employed hierarchical density-based clustering (HDBSCAN) to automatically group semantically similar action terms [31]. This approach effectively handled multilingual synonyms, closely related procedural terms, and minor lexical variations.

For instance, the embeddings enabled accurate clustering of synonyms and closely related terms such as:

– заміна ("replacement"), зняття ("removal"), установка ("installation");

– чистка ("cleaning"), промивка ("flushing"), миття ("washing").

Rule-based and Manual Adjustment

Semantic clustering results were further refined through manual expert review and rule-based adjustments.

Rule-based adjustments utilized explicit semantic rules, for example:

– merge clusters if cosine similarity between cluster centroids exceeded 0.9;

– split clusters containing more than two semantically distinct action groups, identified by keyword-based heuristics (e.g., presence of conflicting action verbs such as "replacement" vs "diagnostics").

Domain experts reviewed cluster coherence, merging overly fragmented clusters, splitting heterogeneous groups, and correcting semantic misalignments.

Construction of the Taxonomy

Clusters were systematically organized into hierarchical branches, forming a structured taxonomy. Synonymous and closely related actions became single taxonomy nodes, while semantically distinct terms formed separate branches. Ambiguities were resolved through manual expert validation. For example, the terms: “регулювання” (“adjustment”) and “налаштування” (“tuning”) became sibling nodes, while clearly distinct concepts like “заміна” (“replacement”) and “діагностика” (“diagnostics”) occupied separate taxonomy branches.

Evaluation of Clustering Quality and Taxonomy Coherence

To quantify taxonomy quality, we conducted an expert-based coherence evaluation. Automotive domain experts manually reviewed a random sample of 120 semantic clusters. Each cluster was evaluated according to two criteria:

– semantic consistency: experts confirmed whether all actions within a cluster accurately represented a single semantic category without irrelevant terms;

– action ambiguity: experts checked whether any action could simultaneously belong to more than one cluster, indicating poor semantic separation.

The coherence score was calculated as the percentage of clusters that fully satisfied both criteria without requiring any modifications. As a result, 111 out of 120 clusters (92.5 %) were deemed semantically coherent without the need for further adjustments, while the remaining 9 clusters (7.5 %) required minor refinements, such as reassignment of individual ambiguous terms to other clusters.

The resulting semantic clusters containing the top-30 most frequent repair actions are illustrated in Fig. 2 (each cluster represented by a unique color). For visualization purposes, high-dimensional embeddings generated by the multilingual-e5-large model were reduced to two dimensions using the Uniform Manifold Approximation and Projection (UMAP) method [32]. The axes on the plot thus represent these two UMAP dimensions, which preserve semantic relationships and distances among clusters. Consequently, actions that are semantically similar appear closer together in the visualization, whereas semantically distinct actions are plotted further apart. This dimensionality reduction enables intuitive visual inspection and verification of the semantic coherence achieved by the clustering process.

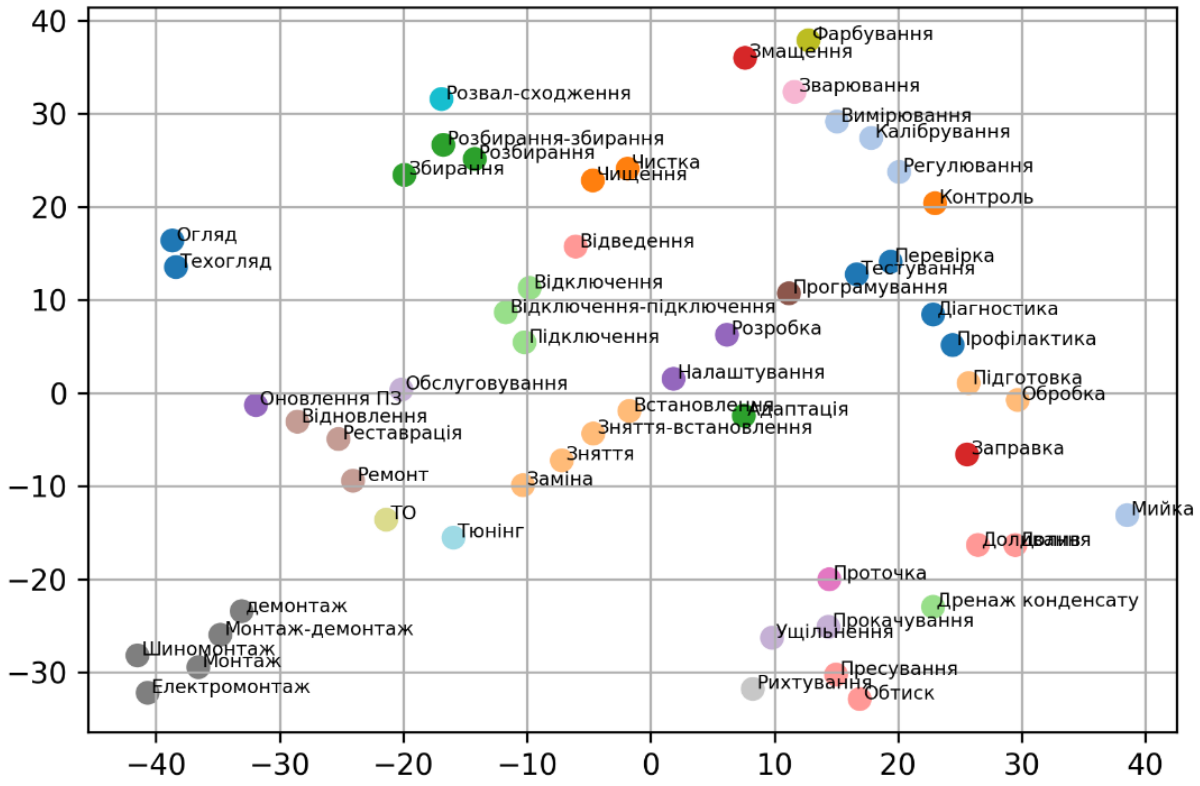


Fig. 2. **Semantic clusters of top-30 automotive repair actions visualized using UMAP dimensionality reduction (multilingual-e5-large embeddings).**

Source: **compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

The resulting taxonomy thus provides a structured, semantically meaningful representation of automotive repair actions, significantly improving analytical capability and system automation.

This hybrid pipeline enabled the successful transition from initial manual annotation through automated classification to the final induction of a structured taxonomy. By combining linguistic preprocessing, semantic classification, and embedding-based clustering, we established a flexible, scalable methodology suitable for handling multilingual, noisy, and dynamically evolving automotive service corpora.

This structured, taxonomy-driven approach provides a clear methodological path toward fully automated, business-ready knowledge organization and operational analytics.

COMPARISON WITH MANUAL ANNOTATION AND EFFICIENCY GAINS

To evaluate the practical effectiveness of the automated taxonomy induction pipeline, we compared it with fully manual annotation, measuring quality, efficiency, and resource implications [33].

Quality and Speed Comparison

Manual annotation by domain experts typically achieves high semantic accuracy; however, it remains extremely labor-intensive and costly when applied to large datasets. For comparison, we evaluated both the quality and speed of our automated pipeline against traditional manual methods. The quality was measured using the previously described cluster coherence metric – percentage of clusters accepted by domain experts without further adjustment. Our automated pipeline achieved a comparable accuracy of 92.5 % coherence after a minimal expert review (sample of 120 clusters).

Regarding speed, annotation time was dramatically reduced due to automation. Manually reviewing and clustering each action from a corpus of 652,929 unique repair names (representative of 4.3 million total repair records) was estimated to take approximately 10 seconds per entry (which involves reading, understanding, assigning semantic labels, and cluster grouping), resulting in approximately 1,800 person-hours for the complete set. Our hybrid pipeline, including preprocessing, automatic classification, clustering, and minimal expert validation, required less than 2 expert-hours initially, demonstrating drastic efficiency gains.

Resource Savings and Efficiency Gains

Automated clustering resulted in substantial savings, reducing annotation costs by approximately 99 %.

Experts from automotive repair garages reported a significant decrease in repetitive manual work, focusing instead on high-value semantic refinement. Furthermore, scalability was greatly improved; enabling continuous, near-real-time taxonomy updating as new data emerged.

Case Studies: Practical Business Benefits

The deployment of automated taxonomy induction in real-world automotive service environments has yielded several tangible business advantages.

1. Comprehensive Repair Statistics

Automated classification and structuring of repair actions enabled the generation of complete and highly detailed statistics on all performed jobs at each service station. Managers gained unprecedented visibility into the frequency, distribution, and types of repairs, supporting data-driven decision-making in staffing, inventory management, and process optimization.

2. Intelligent Chatbot Integration

The structured taxonomy became the foundation for advanced chatbot systems capable of automatically recognizing and interpreting customer requests. These chatbots could, for example, engage customers via messaging platforms, accurately detect the repair or maintenance action required (e.g., “поміняти гальмівні колодки” – “replace brake pads”), and seamlessly initiate appointment scheduling or preliminary cost estimation.

3. Benchmarking Across Departments, Branches, and Countries

With consistent taxonomy applied across multiple service centers, it became possible to conduct benchmarking analyses not only within a single department but also across branches, companies, and even internationally. Service providers could compare repair profiles, operational efficiency, and failure rates, identifying best practices and areas for improvement on a broad scale.

4. Establishment of Industry Standards

The resulting taxonomy, built upon real-world service data and validated by domain experts, provided a practical foundation for the creation of industry standards. This standardization fostered interoperability between information systems, streamlined reporting requirements, and facilitated collaborations across the automotive aftersales sector.

CONCLUSIONS AND RECOMMENDATIONS

This research introduces and validates a hybrid pipeline that successfully automates the transition from simple classification to the structured induction of hierarchical taxonomies, specifically tailored to multilingual automotive repair corpora. By effectively combining rule-based preprocessing, advanced morphological normalization, semantic embedding-based classification, and density-based clustering, the pipeline addresses the persistent challenges of real-world technical text data, including lexical diversity, noise, and the multilingual nature of automotive service records.

The proposed methodology has proven effective at scale, classifying and clustering over 650,000 unique repair actions with minimal manual oversight. Notably, semantic clustering coherence exceeded 92 %, significantly reducing the manual labor traditionally required for such tasks. Furthermore, embedding-based approaches robustly handled the "long tail" of infrequent and emerging action terms, ensuring comprehensive coverage of technical vocabulary.

Practical Recommendations for Industry Deployment

Automotive service organizations and related stakeholders are recommended to:

– adopt standardized taxonomies derived through automated pipelines to enable cross-organizational interoperability, comprehensive benchmarking, and advanced analytics capabilities;

– utilize automated taxonomy structures as a foundation for integrating intelligent digital solutions – such as customer-facing chatbots, appointment schedulers, and automated repair estimators – thus enhancing customer experience and operational efficiency;

– engage domain experts strategically, focusing their efforts primarily on initial seed dictionary development, high-level validation, and periodic taxonomy refinement rather than routine annotation tasks.

Limitations and future improvements

While the proposed pipeline demonstrated strong performance and practical utility, several limitations remain:

– dependency on a manually annotated seed dictionary, requiring domain expertise for initial setup;

– potential sensitivity of semantic clustering to embedding quality and multilingual nuances, particularly for very rare or novel terms;

– computational resource requirements of embedding-based clustering (hdbscan) might limit real-time scalability on large datasets without optimization.

Addressing these limitations through automated seed dictionary generation, enhanced multilingual fine-tuning, and optimized computational approaches is a clear direction for future research.

Future Research Directions

Several promising directions emerge from this study.

1. Automated Expansion of Taxonomies

Future work should investigate fully automated methods for discovering synonyms and taxonomic relationships using large language models (LLMs). Models such as GPT-series or multilingual domain-specific transformers could facilitate ongoing taxonomy growth and adaptation without significant human intervention.

2. Multi-level Taxonomy Development

Extending the pipeline beyond the action level (level 6) to cover higher levels – such as systems, subsystems, units, and components – will further enhance analytics and decision support capabilities. Methodological innovations and further linguistic research are needed to reliably structure these higher taxonomy layers.

3. Adaptation to Other Technical Domains

Applying and testing the hybrid taxonomy-induction pipeline in other technical and multilingual industries (e.g., manufacturing, aerospace maintenance, healthcare) would significantly validate its generalizability and help identify domain-specific adaptations required for optimal performance.

4. Integration with Real-time Data Streams

Future systems should be developed for real-time taxonomy updating, enabling continuous learning from new data streams, predictive analytics, and adaptive management of repair and maintenance processes.

In summary, the research provides a clear, validated framework that bridges the gap between raw, noisy textual data and structured, actionable taxonomies. This hybrid pipeline not only advances the state-of-the-art in automated multilingual taxonomy induction but also sets the stage for broader adoption and digital transformation across industries reliant on complex technical vocabularies.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

REFERENCES

1. Nickerson, R. C., Varshney, U. & Muntermann, J. “A method for taxonomy development and its application in information systems”. *European Journal of Information Systems*, 2013; 22 (3): 336–359, https://www.scopus.com/authid/detail.uri?authorId=22981044500. DOI: https://doi.org/10.1057/ejis.2012.26.

2. Silla, C. & Freitas, A. “A survey of hierarchical classification across different application domains”. *Data Mining and Knowledge Discovery.* 2011; 22 (1): 31–72, https://www.scopus.com/authid/detail.uri?authorId=8707999600. DOI: https://doi.org/10.1007/s10618-010-0175-9.

3. Shen, J. & Han, J. “Taxonomy-Guided classification”. In: *Automated Taxonomy Discovery and Exploration. Synthesis Lectures on Data Mining and Knowledge Discovery. Springer, Cham*. 2022. DOI: https://doi.org/10.1007/978-3-031-11405-2_5.

4. Tahseen, Q. “Taxonomy – the crucial yet misunderstood and disregarded tool for studying biodiversity”. *Journal of Biodiversity & Endangered Species.* 2014; 02 (3), https://www.scopus.com/authid/detail.uri?authorId=16176538400. DOI: https://doi.org/10.4172/2332-2543.1000128.

5. Janssen, A., Donnelly, C. & Shaw T. “A taxonomy for health information systems”. *Journal of Medical Internet Research*. 2024; 26, https://www.scopus.com/authid/detail.uri?authorId=56900648800. DOI: https://doi.org/10.2196/47682.

6. Bach, J., Otten, S. & Sax, E. “A taxonomy and systematic approach for automotive system architectures. from functional chains to functional networks”. *Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2017)*. 2017, https://www.scopus.com/authid/detail.uri?authorId=57190835864.
DOI: https://doi.org/10.5220/0006307600900101.

7. Li, D., Wang, S., He, Q. & Yang, Y. “Cost-effective land cover classification for remote sensing images”. *Journal of Cloud Computing*. 2022; 11 (1), https://www.scopus.com/authid/detail.uri?authorId=57196194973. DOI: https://doi.org/10.1186/s13677-022-00335-0.

8. Olmedilla, M., Martínez-Torres, M. & Toral, S. “The superhit effect and long tail phenomenon in the context of electronic word of mouth”. *Decision Support Systems*. 2019; 125, https://www.scopus.com/authid/detail.uri?authorId=57006716100.
DOI: https://doi.org/10.1016/j.dss.2019.113120.

9. Clever, L., Pohl, J. S., Bossek, J., Kerschke, P. & Trautmann, H. “Process-oriented stream classification pipeline: a literature review”. *Applied Sciences*. 2022; 12 (18), https://www.scopus.com/authid/detail.uri?authorId=57215419723. DOI: https://doi.org/10.3390/app12189094.

10. Smith, V. C., Gonzalez Hernandez, F., Wattanakul, T., et al. “An automated classification pipeline for tables in pharmacokinetic literature”. *Scientific Reports*. 2025; 15, https://www.scopus.com/authid/detail.uri?authorId=57226517670. DOI: https://doi.org/10.1038/s41598-025-94778-5.

11. Sheikhkhoshkar, M., El Haouzi, H., Aubry, A., et al. “From NLP to taxonomy: identifying and classifying key functionality concepts of multi-level project planning and control systems”. *Journal of Information Technology in Construction*. 2024; 29: 1200–1218, https://www.scopus.com/authid/detail.uri?authorId=57201278147. DOI: https://doi.org/10.36680/j.itcon.2024.053.

12. Peng, B., Narayanan, S. & Papadimitriou, C. “On limitations of the transformer architecture”. *Conference On Language Modeling.* 2024. DOI: https://doi.org/10.48550/arXiv.2402.08164.

13. Islam, S., Elmekki, H., Elsebai, A., et al. “A comprehensive survey on applications of transformers for deep learning tasks”. 2023, https://www.scopus.com/authid/detail.uri?authorId=59440450600. DOI: https://doi.org/10.48550/arXiv.2306.07303.

14. Kim, S.-W. & Gil, J.-M. “Research paper classification systems based on TF-IDF and LDA schemes”. *Human-centric Computing and Information Sciences*. 2019; 9 (1), https://www.scopus.com/authid/detail.uri?authorId=57195458965. DOI: https://doi.org/10.1186/s13673-019-0192-7.

15. Rakhmanov, O. “A comparative study on vectorization and classification techniques in sentiment analysis to classify student-lecturer comments”. *Procedia Computer Science*. 2020; 178: 194–204, https://www.scopus.com/authid/detail.uri?authorId=57207757503.
DOI: https://doi.org/10.1016/j.procs.2020.11.021.

16. Sun, J.-W., Bao, J.-Q. & Bu, L.-P. “Text classification algorithm based on TF-IDF and BERT”. *11th International Conference of Information and Communication Technology (ICTech)*. 2022. p. 533–536, https://www.scopus.com/authid/detail.uri?authorId=57193065038.
DOI: https://doi.org/10.1109/ICTech55460.2022.00112.

17. Savelka, J., Agarwal, A., Bogart, C., et al. “Can Generative Pre-trained Transformers (GPT) pass assessments in higher education programming courses?” *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education (ITiCSE 2023)*. 2023; 1: 117–123. https://www.scopus.com/authid/detail.uri?authorId=54783405100.
DOI: https://doi.org/10.1145/3587102.3588792.

18. Mashtalir, S. & Nikolenko, O. “Advancing automotive technical text analysis: a tree-based classification approach for Ukrainian texts”. *Lecture Notes on Data Engineering and Communications Technologies*. 2025; 242: 339–348, https://www.scopus.com/authid/detail.uri?authorId=59739709200. DOI: https://doi.org/10.1007/978-3-031-84228-3_29.

19. Danielkiewicz, R. & Dzieńkowski, M. “Analysis of user experience during interaction with automotive repair workshop websites”. *Journal of Computer Sciences Institute*. 2024; 30: 39–46. DOI: https://doi.org/10.35784/jcsi.5416.

20. Mohammad, A. “Using blockchain for data collection in the automotive industry sector: a literature review”. *Journal of Cybersecurity and Privacy*. 2022; 2 (2): 257–275, https://www.scopus.com/authid/detail.uri?authorId=57768432600. DOI: https://doi.org/10.3390/jcp2020014.

21. Hemphill, T., Longstreet, P. & Banerjee, S. “Automotive repairs, data accessibility, and privacy and security challenges: A stakeholder analysis and proposed policy solutions”. *Technology in Society*. 2022; 71 (3), https://www.scopus.com/authid/detail.uri?authorId=7004192032.
DOI: https://doi.org/10.1016/j.techsoc.2022.102090.

22. Mashtalir, S. V. & Nikolenko, O. V. “Optimizing hierarchical classifiers with parameter tuning and confidence scoring”. *Herald of Advanced Information Technology*. 2024; 7 (3): 231–242. https://www.scopus.com/authid/detail.uri?authorId=59739709200. DOI: https://doi.org/10.15276/hait.07.2024.15.

23. Yang, F. Li, X., Liu, Q., Li, X. & Li, Z. “Learning-based hierarchical decision-making framework for automatic driving in incompletely connected traffic scenarios”. *Sensors*. 2024; 24 (8): 2592, https://www.scopus.com/authid/detail.uri?authorId=57716777200. DOI: https://doi.org/10.3390/s24082592.

24. Mashtalir, S. & Nikolenko, O. “Data preprocessing and tokenization techniques for technical Ukrainian texts”. *Applied Aspects of Information Technology*. 2023; 6 (3): 318–326. https://www.scopus.com/authid/detail.uri?authorId=36183980100. DOI: https://doi.org/10.15276/aait.06.2023.22.

25. Rysin, A. & Starko, V. “Large electronic dictionary of Ukrainian (VESUM)”. Web version 6.6.9. 2005-2025. – Available from: https://vesum.nlp.net.ua. – [Accessed: Apr, 2024].

26. Tan, P.-N., Steinbach, M. & Kumar, V. “Introduction to data mining”. *2nd ed. Published by Pearson Education Limited, Harlow*. 2019.

27. Enevoldsen, K., Chung, I., Kerboua, I., et al. “MMTEB: Massive Multilingual Text Embedding Benchmark”. 2025, https://www.scopus.com/authid/detail.uri?authorId=57222067234. DOI: https://doi.org/10.48550/arXiv.2502.13595 .

28. Vaswani, A., Shazeer, N., Parmar, N., et al. “Attention is all you need”. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. 2017, https://www.scopus.com/authid/detail.uri?authorId=55147219200 .

29. Wang, L., Yang, N., Huang, X., et al. “Multilingual E5 Text Embeddings: A technical report”. 2024. DOI: https://doi.org/10.48550/arXiv.2402.05672.

30. Sokolova, M. & Lapalme, G. “A systematic analysis of performance measures for classification tasks”. *Information Processing & Management*. 2009; 45 (4): 427–437, https://www.scopus.com/authid/detail.uri?authorId=57202694441.
DOI: https://doi.org/10.1016/j.ipm.2009.03.002.

31. McInnes, L., Healy, J. & Astels, S. “hdbscan: Hierarchical density based clustering”. *The Journal of Open Source Software*. 2017; 2 (11), https://www.scopus.com/authid/detail.uri?authorId=57201253315. DOI: https://doi.org/10.21105/joss.00205.

32. Härdle, W., Simar, L. & Fengler, M. “Uniform Manifold Approximation and Projection”. *Applied Multivariate Statistical Analysis*. 2024. DOI: https://doi.org/10.1007/978-3-031-63833-6_23.

33. Zschech, P. “Beyond descriptive taxonomies in data analytics: a systematic evaluation approach for data-driven method pipelines”. *Inf Syst E-Bus Manage.* 2023; 21 (1): 193–227, https://www.scopus.com/authid/detail.uri?authorId=56436770700. DOI: https://doi.org/10.1007/s10257-022-00577-0.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article
Author Sergii V. Mashtalir is a member of the Editorial Board of this journal. This role had no influence on the peer review process or editorial decision regarding this manuscript

Received 11.03.2025
Received after revision 10.06.2025
Accepted 19.06.2025

DOI: https://doi.org/10.15276/hait.08.2025.9
УДК 004.91

Від класифікації до таксономії: автоматизоване структурування назв робіт з ремонту автомобілів у багатомовних корпусах

Машталір Сергій Володимирович[1)]
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
Ніколенко Олександр Володимирович [2)]
ORCID: https://orcid.org/0000-0002-6422-7824; oleksandr.nikolenko@uzhnu.edu.ua. Scopus Author ID: 59739709200
[1)] Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна
[2)] Ужгородській національний університет, вул. Університетська, 14. Ужгород, 88000, Україна

АНОТАЦІЯ

У цьому дослідженні запропоновано й ретельно перевірено гібридний п’ятиетапний підхід до обробки природньої мови (Natural Language Processing) , який перетворює неструктуровані двомовні тексти про роботи з наряд-замовлень для ремонту автомобілів на багаторівневу ієрархічну класифікацію робіт. Підхід ліквідує розрив між класичною класифікацією за ключовими словами та бізнес-орієнтованою організацією даних. Враховуючи обмеження як традиційних, так і сучасних NLP-методів у технічних, зашумлених і галузево-специфічних датасетах, запропонована методологія об’єднує: розвинену лематизацію, ручне створення словника-ядра, семантичну фільтрацію, класифікацію на основі трансформерів і кластеризацію за векторними представленнями. Спираючись на вдосконалену українську лематизацію, динамічну семантичну фільтрацію, реченнєві вкладення та кластеризацію на основі густини, запропонований алгоритм послідовно нейтралізує шум, багатомовність і «довгий хвіст», притаманні реальним даним по автомобільним ремонтам. Підхід був випробуваний на корпусі з понад 4,3 млн сервісних записів. Він досяг понад 92 % когерентності кластерів, потребуючи лише мінімальний обсяг ручної анотації. Сформовані стандартні довідники відкривають чотири безпосередні переваги для бізнесу: аналітику та порівняння ремонтів на рівні філій, мереж і брендів; чат-боти з розумінням запитів і намірів для точного визначення заявок і автоматизованого розрахунку кошторисів; оптимізацію запасів і робочого часу завдяки деталізованій статистиці робіт; практичну стандартизацію номенклатури ремонтів, яка сприяє обміну даними в межах галузі. Показано, що поєднання мінімального експертного вкладу із сучасними техніками векторних подань і кластеризацією на основі густини, дає змогу автоматизувати створення довідників у промислових масштабах. Це встановлює новий орієнтир для проєктів цифрової трансформації, що залежать від точної структуризації даних на основі зашумлених технічних виразів.

Ключові слова: обробка природної мови; індукція таксономій; семантична кластеризація; машинне навчання; аналіз даних; прикладні інтелектуальні системи; автоматизація, керована даними; організація знань; автоматизація бізнес-процесів.

ABOUT THE AUTHORS



Sergii V. Mashtalir – Doctor of Engineering Science, professor, Informatics Department. Kharkiv National University of Radio Electronics, 14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
Research field: Image and video processing; data analysis

Машталір Сергій Володимирович – доктор технічних наук, професор кафедри Інформатики. Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна



Oleksandr V. Nikolenko – Specialist on Applied Mathematics. PhD student. Uzhhorod National University, 14, University Str. Uzhhorod, 88000, Ukraine
ORCID: https://orcid.org/0000-0002-6422-7824: oleksandr.nikolenko@uzhnu.edu.ua. Scopus Author ID: 59739709200
Research field: Natural language processing; big data; machine learning

Ніколенко Олександр Володимирович – спеціаліст за спеціальністю «Прикладна математика». Здобувач ступеня доктора філософії. Ужгородський національний університет», вул. Університетська, 14. Ужгород, 88000, Україна