

Computational intelligence technology for Ukrainian language textual content processing based on big data analysis, NLP and machine learning

Victoria Vysotska

Lviv Polytechnic National University, Stepan Bandera 12, 79013 Lviv, Ukraine

Abstract

The work aims to develop models, methods, and means of analysis and synthesis of computer linguistic systems (CLS) based on new and improved methods of processing Ukrainian-language textual content to solve natural language processing problems (NLP). The scientific novelty of the obtained results lies in solving an important scientific and applied problem of analysis and synthesis of CLS for solving various tasks of processing Ukrainian-language textual content based on developing new and improving known models, methods and means of NLP. The following new scientific results were obtained: – A model of intellectual analysis of the text flow, which, unlike the existing one, is based on the processing information resources, NLP and machine learning, which the typical structures of content integration, management and support modules; – Methods of adapted processing information resources for processing Ukrainian-language text and take into account the needs of the permanent target audience based on the analysis of the history of the target audience's activity on the CLS web resource, which made it possible to form a set of metrics and indicators of the effectiveness of the CLS functioning for the various NLP tasks solution; – A model of linguistic processing of text based on the grapheme, morphological, lexical and syntactic analyses improvement, which, unlike the existing ones, are adapted for processing Ukrainian-language text through regular expressions and machine learning, made it possible to adapt the processes of processing Ukrainian-language text content and increase the accuracy of the obtained results depending from a specific NLP task; – A method of identifying keywords in Ukrainian-language texts based on grapheme and morphological analysis of word bases through regular expressions and N-grams was developed, which made it possible to increase the accuracy of searching for keywords, search for stable word combinations and categorize content; – A method of determining the style of the author of thematic Ukrainian-language text content was developed based on the keywords, stable word combinations, N-grams analysis, which made it possible to determine the stylistic contribution of each of the authors and increase the accuracy of the attribution of a scientific and technical publication; – A method was developed for calculating the degree of verification of the author of a Ukrainian-language text from a set of possible ones based on a comparative analysis of the styles of potential authors, which made it possible to increase the accuracy of classification based on the similarity of style; – Methods of analysis and synthesis of CLS were developed based on the creation of a general typical structure of the text content processing CLS in the Ukrainian language through support for modularity, modelling of the interaction of main processes and components, which made it possible to expand the collection of solutions to various typical tasks of the NLP by implementing typical software of such systems; – NLP methods, which, unlike the existing ones, are implemented on the basis of developed regular expressions of grapheme and morphological analysis of Ukrainian-language text and modified Porter's stemming algorithm as an effective identifying lem affixes for the possibility of demarcating the analysed word, which made it possible to optimize the process and improve the accuracy of Ukrainian words/sentences normalization; – Text tokenization and normalization methods, which, in contrast to the existing ones, use cascades of simple substitutions of developed regular expressions of matching with templates based on production rules, finite automata and the ontological model of the rules of the Ukrainian language syntax.

Keywords

Computer linguistic systems, NLP, Ukrainian-language, textual content, machine learning

CLAW-2024: Computational Intelligence Application Workshop, October 10-12, 2024, Lviv, Ukraine

✉ Victoria.A.Vysotska@lpnu.ua (V. Vysotska)

ORCID ID 0000-0001-6417-3689 (V. Vysotska)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Introduction

The active development of information technologies (IT) is at the intersection of globalization and informatization. The rapid rate of growth of society's informatization is directly related to the rate of development and implementation of computer linguistic systems (CLS), the development of which is based on models and methods of natural language processing (NLP) [1-3]. The complexity of developing models, techniques, and tools of NLP lies in solving non-typical NLP problems and adapting these models, methods, and tools to a specific natural language [4-6]. Each natural language is unique, with its flavour of rules, history, grammar, exceptions, and peculiarities of generating linguistic units for conveying meaning, complicating developing a CLS.

Usually, each successful CLS development project is designed for a specific task (for example, machine translation [7-9], identification of plagiarism/rewriting [10-12], text rubrication [13-14], text attribution analysis [15-21], information retrieval [22-28], referencing/abstracting [29-30], voice assistants [31-33], intelligent chatbots [34-39], etc.) and is both one-time and closed (for example, Amazon Alexa, Google Assistant, Facebook, Voice Mate, Bixby, Siri, Abby Lingvo, Microsoft Cortana, Microsoft Word, Grammarly, Google Translation, PROMT, CuneiForm, Trados, OmegaT, Wordfast, Dragon, IBM via voice, Speereo, Finereader, Tesseract, OCRopus, etc.) without being able to read the content to willing IT professionals/specialists. In rare cases, the developers provide open access to such CLS projects and the opportunity to get acquainted with their structure and content. The development of any NLP application for an arbitrary natural language of more than 7000 languages and dialects is based on studying large textual monolingual/parallel corpora of that language, containing more than hundreds of millions of words and linguistic resources. Only about 20 natural languages (English, Chinese, Western European languages, Japanese, etc.) are the results of research on such corpora known, making it possible to develop CLS of various complexity for these languages. Unfortunately, in modern realities, the Ukrainian language is considered in the international scientific community to be an exotic language with a low resource index, i.e., it does not have enough educational, research and processed data to develop modern applied applications of NLP. Such applied applications are used to build CLS in cyber security (detection of fakes and propaganda, so-called trolls/bots in social networks), sociology (analysis of the dynamics of changes in public opinion on thematic issues), philology (automatic research of large data sets of various thematic orientations and different periods), psychology (analysis of the psychological portrait of a person, identification of post-traumatic stress disorder of participants in hostilities or occupation), national security (information warfare), jurisprudence (criminology and court case), social communications (analysis of community posts in social networks) and other important branches of modern Ukraine. The above determines the relevance of the topic of the dissertation research.

Scientific research by N. Chomsky, V.M. Glushkov, A.V. Hladkoy, D.V. Lande, V.A. Shyrokov, N.V. Sharonova, N.F. Khairova, O.V. Bisikalo, S.N. Buk, N.P. Darchuk, Z.V. Partyka, A.V. Anisimova, Yu.D. Apresyan, O.O. Marchenko, I.M. Kulchytskyi, A.O. Nikonenko, M. Gross, A. Lanten, V.H. Yngve, S. Sharoff, Yu.A. Schrader, D. Jurafsky, B. Bengfort, J.H. Martin, L. Tesniere, T. Ojeda, P.M. Postal, D.G. Hays, T.A. van Dijk, S. Marcus, J. Lyons, L.W. Tosh, Y. Bar-Hillel, D.G. Bobrow, G. Lakoff, R. Bilbro, N. Kotsyba, A.Yu. Berko, Yu.M. Shcherbyna, V.Yu. Velychko, V.F. Starko and many others make it possible to understand the basic principles of linguistic processing of the text depending on the features of a specific natural language. More than 80% of such studies concern the processing of English-language texts. There are fewer studies on Slavic languages, particularly the low-resource Ukrainian language. In particular, there are no publications regarding the development recommendations, functional requirements, general structure, or typical architecture of the CLS for processing Ukrainian-language textual content. Directly applying the English language's models, methods, algorithms, and IT processing to Ukrainian-language textual content does not yield positive results. Already at the level of morphological analysis, a significant conflict arises between the methods developed for the English-language text and their use for the Ukrainian-language text. For example, for a simple Porter algorithm (stemming) without appropriate modification, it is not correct to separate the base of the word from the inflexion, which leads to inaccurate identification of key

phrases, which, in turn, affects the solution of any NLP problem where it is necessary to quickly identify set of keywords (categorization, search, annotation, etc.). Determining the main features and processes of linguistic analysis of Ukrainian-language texts will significantly facilitate the stages of processing the text flow of information, such as integration, support and content management. In turn, the adaptation of the processes of intellectual analysis of text content with the identification of functional requirements for the relevant modules of the CLS will lead to the possibility of developing its typical architecture based on the principle of modularity (adding components depending on the content of the NLP task and the purpose of the CLS).

The above testifies to the relevance of research in solving the significant scientific and applied problem of analysis and synthesis of CLS for solving various tasks of processing Ukrainian-language textual content, which will make it possible to increase the level of resourcefulness of the natural Ukrainian language based on the development of new and improvement of known models, methods and means of NLP.

The work aims to develop models, methods, and means of analysis and synthesis of computer linguistic systems based on new and improved known methods of processing Ukrainian-language textual content to solve problems of natural language processing. The purpose of the work is to determine the need to perform such tasks:

1. To analyse the specifics of the construction of the CLS by systematizing the processes of their implementation and functioning, which will provide an opportunity to distinguish a class of systems whose functional properties allow to perform a quantitative assessment of the expected effects of the implementation of a typical CLS of processing Ukrainian-language textual content for solving various tasks of the NLP;
2. To develop information technology for the construction of CLS for the processing of Ukrainian-language text, which will make it possible to determine their basic structure, functional requirements, the sequence of setting and training the system, and general design principles;
3. To offer IT processing of information resources as integration, management and support of Ukrainian-language content based on the improvement of linguistic analysis of text content for the development of metrics for evaluating the effectiveness of the functioning of the CLS for solving various tasks of the NLP;
4. To develop methods of processing Ukrainian-language textual content for solving various problems of NLP to increase the accuracy of the obtained results;
5. To develop methods and means of intellectual analysis of textual content to increase the efficiency of solving various tasks of NLP;
6. Create software modules for processing Ukrainian-language textual content for solving various tasks of NLP and conducting experiments;
7. To test the obtained results by building and implementing applied CLS to process Ukrainian-language textual content.

The object of research is the processes of analysis and synthesis of computer linguistic systems for processing Ukrainian-language textual content.

The research subject is models, methods, and means of processing Ukrainian-language textual content to solve various problems of NLP.

The following research methods were used to achieve the goal: the theory of formal grammars and automata, the theory of sets, the theory of data and knowledge models, the theory of probability and mathematical statistics, the theory of models, algorithms, and logical-linguistic numbers, information theory, graph theory, and knowledge presentation methods for modelling the processes of processing Ukrainian-language textual content and developing machine learning modules; models and methods of processing and analysing textual content for the implementation of the processes of solving various problems of NLP; methods of object-oriented and system analysis and design - for design and development of CLS; the theory of relational databases, methods of artificial intelligence,

object-oriented programming - for the software implementation of the Ukrainian-language textual content processing system for the solution of various NLP tasks. The practical significance of the obtained results lies in the fact that they can be used to build applied CLS for processing Ukrainian-language textual content. In particular, the following results are practically valuable:

- The application of the method of identification of persistent word combinations in the identification of keywords in Ukrainian-language scientific texts of a technical profile allows an increase in the accuracy of the search for keywords by 6-9% and highlights thematic terms from the text for further classification of the publication;
- Development of a formal approach to the design of a content monitoring module for identifying keywords in Ukrainian-language texts based on web data mining, NLP and linguistic analysis of defined words of text content, which made it possible to develop the general structure of typical CLS and increase the effectiveness of CLS functioning by 6-9% depending on the solution of a specific NLP problem;
- The application of the method of calculating the degree of verification of the author of the Ukrainian-language text based on the analysis of the styles of potential authors made it possible to increase the accuracy of identification by 6-12% and carry out the decomposition of the method through the study of stylistic coefficients such as the coherence of speech, the degree of syntactic complexity, linguistic diversity, indices of concentration and exclusivity of the text;
- Development of a content monitoring module to identify a potential author of a text from a set of possible ones based on a comparison of the results of the analysis of a template author's text with the researched one to reduce the volume of the corresponding set to [9;34]% of the total number of project participants, depending on the subject and the time range of scientific writing - technical publications, as well as the frequency of publications of this author in this period on a specific topic;
- Experimental testing of the method of identifying the author's style in Ukrainian-language texts based on web data mining and linguistic analysis of defined stop words allows the selection of content potentially similar in style from a set of potential author's publications.

2. Related works

Determining the main processes and features of the linguistic analysis of Ukrainian-language texts will significantly facilitate the stages of processing the text flow of content such as integration, support and content management (Fig. 1). Adaptation of the processes of intellectual analysis of text content with the identification of functional requirements for the relevant modules of the CLS will lead to the possibility of developing a typical structure of similar systems based on the principle of modularity (adding components depending on the content of the NLP task and the purpose of the CLS). The application of the specified IT/methods/models in the typical structure of the CLS, adapted for any process of processing Ukrainian-language textual content, is a necessary prerequisite for the successful implementation of the CLS project for solving a specific task of the NLP, which requires the use of an appropriate set of standard libraries, utilities and software with open source, which will solve specialized functions of the project according to the needs of the end user. The state of the CLS is determined by the tuple of the main properties at a specific moment in time or the activity of the corresponding NLP process: $s_i = (p_{i1}, p_{i2}, \dots, p_{im})$, $i = \overline{1, n}$, where s_i is the corresponding i -th state at a specific moment in time t_i from the set with power $|S|=n$, p_{ij} is the corresponding ij -th property of the state from the set with power $|P|=m$, which determines the behaviour of the CLS as $p_j = (r_{ij1}, r_{ij2}, \dots, r_{ijv})$, $j = \overline{1, m}$, where r_{ijk} is the corresponding parameter of the specific property p_{ij} for the state s_i . For any CLS, the state s_i is one of the NLP processes, for example, the identification of keywords and/or stable phrases for the next state s_{i+1} of the system as a rubric of a text array of data. Accordingly, the properties of the state s_i are morphological p_{i1} , lexical p_{i2} and syntactic p_{i3} .

Some NLP tasks may have semantic ones, etc. Then, for the property p_j , a set of parameters is determined for the corresponding text analysis, depending on the specific task of NLP [40-50]. According to these parameters, the strategy of the CLS operation at the moment of time t_i is specified for:

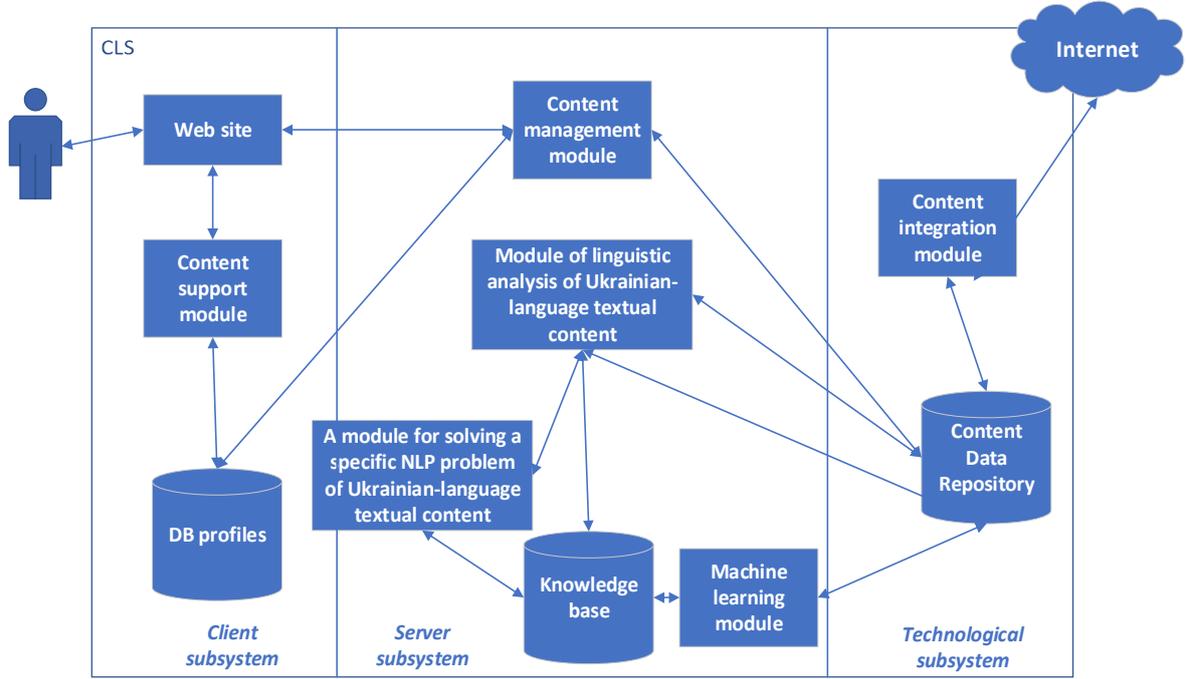


Figure 1: Generalized structure of the computer linguistic system

- parameters of the morphological property p_{i1} are N-grams and morphemes: roots r_{i11} , endings r_{i12} , affixes r_{i13} ; grammatical categories of different parts of speech r_{i14} , word length r_{i15} , word placement in a sentence r_{i16} , number of syllables in a word r_{i17} , number of word contents r_{i18} , ratio of consonants and vowels r_{i19} , etc.;
- the parameters of the lexical property p_{i2} are the location of the sentence in the text r_{i21} , the location of the word in the sentence r_{i22} , the weight of the word r_{i23} , the weight of the sentence r_{i24} , the base of the word r_{i25} , the inflexion of the word r_{i26} , etc.;
- parameters of the syntactic property p_{i3} are the depth of the word in the dependency tree of the sentence r_{i31} , the location of the word in the sentence r_{i32} , the number of contents of the word r_{i33} , the number of words per sentence r_{i34} , the number of words r_{i35} and sentences r_{i36} , whether the word is a capital letter r_{i37} / with a hyphen r_{i38} / compound r_{i39} , etc.;
- parameters of the semantic property p_{i4} are the number of word content r_{i41} , the depth of the word in the dependency tree r_{i42} , the size of paragraphs r_{i43} , the placement of paragraphs r_{i44} , etc.

Depending on the tuple $p_j \in s_i$, the behaviour of the CLS is determined, that is, the implementation of a set of rules (activation of actions or events) for implementing a specific NLP process depending on the input text data. Accordingly, the event o_l is the change of one property to another $p_{ij} \rightarrow p_{ik}$ or $o_l: p_i \rightarrow p_j$ according to the fulfilment of certain conditions U for the input analyzed text X and the intermediate processed text $C: p_i = o_l(p_j, U, X, C)$. Action d_g is the process of activation of an event o_l by another event o_v in CLS: $C' = d_g(o_l \circ o_v)$. The more complex the language (morphology, syntax, etc.), the more difficult it is to process the corresponding texts in natural language. In addition, for such low-resource languages as Ukrainian, there are no standardized rules and dictionaries for processing texts in natural language to solve the relevant tasks of NLP. Many

scientific linguistic schools and IT specialists are working on creating Ukrainian dictionaries, text corpora and rules for processing Ukrainian texts. However, these are usually linguists and philologists unfamiliar with the features of specific modern tools, such as programming languages, ML methods, big data analysis, etc. There is a colossal gap between the research results of philologists and applied linguists, on the one hand, and IT specialists, on the other, for developing Ukrainian-language tests. Today, quite a few, such as Ukrainian, have been implemented for general access to NLP tools.

3. Material and methods

The developed typical structure of S_{wtm} CLS consists of modules for solving a specific task of NLP M_{dis} , content support M_{dmr} , content integration M_{dcp} , content management M_{dvm} , linguistic M_{lat} and intelligent analysis of textual content flows (IATCF) M_{was} [48]:

$$S_{wtm} = \langle M_{dis}, M_{dmr}, M_{dcp}, M_{dvm}, M_{lat}, M_{was} \rangle. \quad (1)$$

Accordingly, the solution module of a specific NLP problem M_{dis} :

$$M_{dis} = \langle N_{wvr}, S_{gcc}, S_{gco}, S_{gcv}, S_{gro}, P_{wnv}, I_{wnv} \rangle, \quad (2)$$

where S_{gcc} is the average conversion rate, S_{gco} is the average cost of orders, S_{gcv} is the average cost or utility of the purpose of the visit, S_{gro} is the average P_{ROI} or the average return on investment, P_{wiv} is the percentage (%) of profit from new visitors, I_{wnv} is the new buyers/customers index at the first visit.

The presence of the M_{dmr} text content support module reduces costs for moderators/analysts who collect/analyze statistical data on the dynamics of the CLS functioning, the activity of the permanent target audience as a reaction to website content changes, and the formation of rules for the analysis of user information portraits and thematic content plots:

$$M_{dmr} = \langle I_{gyk}, K_{gvb}, P_{wap}, P_{wvk}, S_{grk}, I_{gck}, P_{wck}, P_{wvk}, K_{wcz}, P_{wvz} \rangle, \quad (3)$$

where I_{gyk} is the advertising quality index; K_{gvb} is a brand recognition factor; P_{wap} and P_{wvk} are % of new/repeated customers and users; S_{grk} is average P_{ROI} by type of advertising; I_{gck} and P_{wck} are index and % conversion of goals by type of advertising; P_{wvk} and P_{wvz} are % of visits by type of media advertisement; K_{wcz} is the conversion rate of goals by type of means.

$$I_{gyk}(w) = \frac{P_{wcv}(w)}{P_{wvk}(w)}, K_{gvb} = \frac{N_{ubq} + N_{utv}}{N_{uaq} + N_{utv}}, \quad (4)$$

where $P_{wvk}(w)$ is a function for determining % of visits from advertisement w ; $P_{wcv}(w)$ is a function for determining % conversion of goals for visits from w ; $I_{gyk}(w)$ is a function for determining the index of advertising quality w ; N_{uaq} is the total number of user queries of intellectual and informational search (IIS) by keywords; N_{utv} is the number of direct visits to the website; N_{ubq} is the number of IIS requests with brand name.

The presence of the M_{dcp} text content integration module reduces the costs of CLS moderators and content authors, automating/implementing some of their work/functions such as content collection from several different reliable sources, its recognition, filtering, saving, formatting, analysis, annotation, classification, etc.:

$$M_{dcp} = \langle P_{glt}, P_{gst}, P_{ght}, K_{gvb}, K_{uzv}, P_{uav}, P_{uzv}, S_{gnc}, P_{wvv}, S_{gpp}, S_{gtp} \rangle, \quad (5)$$

where P_{glt} , P_{gst} and P_{ght} are % of repeat visits of the user from the previous visit $>t_2$, within $[t_1; t_2]$ when $t_1 < t_2$ and $<t_1$ days, respectively; K_{gvb} is a brand recognition factor; P_{uav} and P_{uzv} are % of new/repeated visitors and interest; S_{gnc} is the average number of clicks on advertising for N_{wvr}

visits; P_{wvv} is the bounce rate for one web page; S_{gpp} is the average number of web page views per visit; S_{gtp} is the average length of stay on the web page.

$$P_{wvv} = \frac{N_{vnp}}{N_{inp}}, \quad S_{gnc} = \frac{N_{wcr}}{N_{wav}} \cdot N_{wvr}, \quad K_{uzv} = \frac{N_{wad}}{N_{wav}}, \quad P_{uzv} = \frac{N_{wzv}}{N_{wvk}}. \quad (6)$$

where N_{inp} is the number of direct web page visits; N_{vnp} is the number of one-page visits to a web page; N_{wvr} is the number of visits for analysis; N_{wav} is the total number of visits; N_{wcr} is the average number of clicks on advertising; N_{wad} is the total number of actions on the page; N_{wvk} and N_{wzv} are the total number of all and interested users.

The presence of a text content management module reduces costs for moderators/administrators who update the website and create rules for caching/searching popular information blocks:

$$M_{dvm} = \langle K_{wis}, P_{wep}, P_{gum}, P_{gup}, P_{gur}, P_{gus}, P_{gub}, P_{gul}, P_{wep}, K_{wdu}, S_{wdu} \rangle, \quad (7)$$

where K_{wis} is an indicator of internal IIS; P_{wep} is % edition of the page with an error; P_{gum} and P_{gup} are % of mobile users with a high-speed Internet connection; P_{gur} and P_{gus} are % of users with low/medium/high display resolution and with a specific operating system; P_{gub} and P_{gul} are % of users with a specific browser and with English and/or Ukrainian language support; K_{wdu} is an indicator of the number of users, views and page visits. The S_{wdu} indicator is the base of the content management module:

$$S_{wdu} = \langle N_{svt}, N_{sut}, N_{spt}, N_{spv} \rangle, \quad (8)$$

where N_{spv} and N_{spt} are the average number of page views per visit and for a specific time Δt ; N_{sut} is the average number of unique users for a specific time Δt ; N_{svt} is the average number of visits for a specific time Δt . The indicator of internal search on the site:

$$K_{wis} = \langle N_{nns}, P_{uts}, P_{ksp}, P_{bus}, P_{cus}, P_{pop}, P_{ucs}, S_{vrs}, P_{uos},$$

$$P_{uns}, P_{unr}, P_{uur}, S_{nup}, T_{svs}, P_{uis}, P_{nrp}, K_{wps} \rangle,$$

where N_{nns} is the number of zero search results; P_{uts} and P_{ksp} are % of users who were on the page for $> t$ time and viewed $> k$ pages after the search; P_{bus} and P_{cus} are % of purchases made and % of buyers among users using search; P_{pop} is % of rejections after visiting one page as a search result; P_{ucs} is % conversion from users using search; P_{unr} and P_{uur} are % of users who do not use and use search; S_{nup} is the average number of pages viewed by visitors after a search; T_{svs} is the average time spent on the site for a visit after a search; P_{uns} and P_{uos} are % of visitors who conduct several searches during the visit and who left the site after viewing the search results; S_{vrs} is the average number of search results; P_{uis} is % of visits with search; P_{nrp} is % of zero search results, in particular,

$$P_{wep} = \frac{N_{wep}}{N_{wpp}}, \quad P_{nrp} = \frac{N_{nps}}{N_{vps}}, \quad K_{wps} = \frac{N_{wsv}}{N_{wns}}, \quad (9)$$

where N_{wpp} , N_{wep} and N_{vps} are the number of all viewed pages issued with an error and viewed pages with a search, respectively; N_{nps} is the number of zero search results; N_{wns} and N_{wsv} is visits without search and with search.

The presence of a module for intellectual analysis of text streams of content reduces the time/costs/personnel/resources for the timely and prompt acquisition of relevant, unique, current content, which leads to an increase in the volume of the target audience of CLS, in particular, contributes to the growth of the economic effect of the implementation:

$$M_{was} = \langle S_{wcc}, S_{wtv}, S_{wnv}, P_{wuv}, P_{wnv} \rangle, \quad (10)$$

where S_{wcc} is the average conversion rate; S_{wtv} is the average length of visit; S_{wnv} is the average number of views per visit; P_{wuv} is % of unique customers/visitors/users; P_{wnv} is % of new website customers.

According to the tracking of K_{as} events and interaction with the K_{du} site, they analyze:

$$K_{usa} = \alpha(K_{wdu}, K_{was}) = \langle P_{vcu}, P_{sau}, P_{siu}, I_{wdx} \rangle, I_{wdx} = \frac{R_{wcv} + R_{wec}}{N_{upv}}, \quad (11)$$

where P_{siu} is % interaction with the site (for example, commenting, voting, registration, authorization, subscription, etc.); P_{sau} is % of users who activate various events (for example, clicking on an ad, starting a function, pausing, etc.); P_{vcu} is % of users interacting with different types of content presentation (viewing the next communication, panning, zooming, etc.); I_{wdx} is the value of the measure of usefulness, respectively, of the page/site/CLS/content; N_{upv} is the number of unique page views; R_{wec} is profit from e-business; R_{wcv} is the value of the utility measure of user visits (based on transactions) and the purpose of user visits (based on the utility of goals).

Analysis of success/effectiveness/operational search on the site:

$$K_{iip} = \langle P_{wuv}, R_{ecc}, S_{wcv}, P_{wip}, P_{wcv}, N_{wvt}, R_{wcv}, R_{wec}, N_{wtr}, N_{wcv}, I_{ssp} \rangle, \quad (12)$$

where P_{wuv} is the value of the usefulness of visiting P_{wuv} site/page; R_{ecc} is conversion rating in e-business for CLS corresponding to the NLP task; S_{wcv} is the value of average utility; P_{wip} is the value of e-business profit for the CLS of the corresponding NLP task; P_{cv} is the value of the achieved conversion of visits to the site/page of the CLS:

$$P_{wuv} = \frac{R_{wcv} + R_{wec}}{N_{wvt}}, R_{ecc} = \frac{N_{wtr}}{N_{wvt}} \cdot 100\%, S_{wcv} = \frac{R_{wcv} + R_{wec}}{N_{wcv} + N_{wtr}}, P_{wip} = R_{wcv} + R_{wec}, P_{wcv} = \frac{N_{wcv}}{N_{wvt}} \cdot 100\%,$$

where N_{wvt} is the number of visits; R_{wec} is the usefulness of e-business; R_{wcv} is the utility of the goal; N_{wtr} is the number of transactions; N_{wcv} is the number of conversions.

To attract new visitors and increase the volume of the permanent target audience, the calculation of the impact on the income of the IIS on the site is used I_{ssp} :

$$I_{ssp} = (R_{ssv} - R_{snv}) \cdot N_{ssv}, \quad (13)$$

where N_{ssv} is the number of visits from the IIS; R_{snv} and R_{ssv} are the utility of visits without and with IIS.

The topic of a set of keywords is one of the main indicators of IIS for identifying the specific content of a page. Optimize investment for sets of keywords that increase conversion values. The return on investment value (P_{ROI}) must be positive ($N_{Inc} > N_{Exp}$), i.e.:

$$P_{ROI} = \frac{N_{Inc} - N_{Exp}}{N_{Exp}} \cdot 100\% > 0, P_{ROIvp} = \frac{(N_{Inc} \cdot A_{Inc})/100 - N_{Exp}}{N_{Exp}} \cdot 100\%, \quad (14)$$

where N_{Exp} is expenses; N_{Inc} is profit; A_{Inc} is the amount of profit. Then they find how much $>q\%$ of funds can be spent on a specific keyword in advertising without the risk of getting $P_{ROI} < 0$. To calculate the amount of funds for attracting users, use:

$$C_{amax} = \frac{\frac{N_{Inc} \cdot A_{Inc}}{100}}{\frac{P_{ROIvp}}{100} + 1}, C_{cmax} = C_{amax} \cdot \frac{R_{ecc}}{100}. \quad (15)$$

The method of determining the effectiveness/quality of the CLS site for solving the NLP problem:
Stage 1. Formulation and identification of usefulness according to the goals of the target audience according to the input data from the tuple X .

Stage 2. Activation of reports of the operation of the CLS from the tuple Y of the initial data:

Step 1. Define an unlimited number of goals (≈ 4 goals for each target audience profile).

Step 2. Identify the optimal volume of visits/time of the end user/customer for a successful conversion.

- Step 3. Analyse the volume of the contribution of each goal to the total profit.
 Step 4. Combine goals by categories/directions/species.
 Step 5. Form separate sets of transactions as appropriate for the purposes.
 Stage 3. Support various marketing campaigns/customers through M_{dmr} .
 Stage 4. Support for processing the service content of the site with the M_{dvm} module.
 Stage 5. Updating the profiles of the target audience according to feedback support through the M_{dmr} module, and analyzing user actions through the M_{dvm} module.
 Stage 6. Integrating content from different sources through M_{dcp} according to the achieved goals and processing it through the M_{was} module.
 Stage 7. Periodic checks are performed to see whether the goals are being achieved and whether the profit is growing according to the goals. If it subsides, go to stage 1. Otherwise, go to stage 2.

A classified list of the input stream of content X with a set of relevant properties demarcates project participants through their typification and restriction of access rights depending on the content: regular users, potential visitors, linguists, statistical analysts, administrators, content/rules moderators, authors of unique content, information resource as content source etc. The typed structure of the content input stream template with a set of relevant properties helps to define the main functional requirements for the site/CLS and its typical structure and delineate the non-functional capabilities, classify the sources, calculate the frequencies and the corresponding restrictions/conditions of integration from the usual source:

$$X = \langle X_a, X_s, X_q, X_f, X_s, X_w, X_b, X_d, X_k, X_v, X_u, X_r, X_t, X_o \rangle, \quad (16)$$

where X_a is URL addresses of sources for databases (DB) of CLS filters; X_s is content as a result of integration from different X_a sources according to a predetermined list of URLs without a predetermined structure according to relevant thematic requests; X_q is thematic requests of visitors/users of the CLS site in the form of a set of keywords or persistent phrases; X_f is actual data of permanent users/profiles and a set of rules of permitted actions within the corresponding type of user of the CLS; X_s is statistical data of actions/ events/ phenomena of the subjects/objects of the CLS for the solution of the corresponding NLP task and the rules for collecting/saving/analysing statistics in specific time intervals of the CLS operation; X_w is statistical data on the functioning of the CLS; X_b is contents of the DB/DS of content/rules/filters/annotations, etc. of the CLS; X_d is different types of linguistic dictionaries depending on the purpose of the CLS for solving a specific NLP problem; X_k is a set of personalized/anonymous reviews and comments of users to the relevant content of CLS; X_v is a tuple of the results of personalized/anonymous votes of regular/potential users regarding the content of CLS; X_u is statistical personalized individual actions of users of the CLS; X_r is set of external/internal advertising of thematic content; X_r is thematic stickers of information content (exchange rates, announcements, digests, weather, anecdotes, horoscope, etc.); X_o is a tuple of options for setting up and changing the CLS/site configurations.

Filling the tuple of the output data stream Y according to the purpose of the CLS for solving a specific NLP problem directly depends on the content of the input classified stream of content X with a predetermined set of properties depending on the interaction with the site of the corresponding types of project participants:

$$Y = \langle Y_c, Y_q, Y_a, Y_v, Y_s, Y_p, Y_t, Y_r, Y_o, Y_k \rangle, \quad (17)$$

where Y_c is text content as an information product or the result of providing an appropriate information service for solving a specific NLP task on the CLS website; Y_q is a set of meaningfully generated/cached pages as a result of thematic requests/IIS of users/visitors of the CLS site; Y_a is annotations/digests/abstracts on textual thematic content; Y_v is a tuple of statistics of user/visitor interaction with the site; Y_s is a tuple of the content of the profiles of regular users of the CLS according to the personalized statistics Y_p for the corresponding generation of an individual portrait of the user/audience at certain time intervals; Y_p is a tuple of meaningful recommended site content,

personalized for a specific regular user according to the profile/actions/interaction with the CLS in certain time intervals; Y_t is a set of content topics/headings with the possibility of renewal according to the results of the latest IIS/requests from regular site users; Y_o is a scheme of interrelationships of textual thematic content according to the appropriate classification (current, relevant, author's, outdated, popular, similar, last-viewed, often-viewed, consecutively by a certain most viewed, longer viewed, most viewed from search engines or internal IIS, viewed by a typical group of users, etc.); Y_r is the set of content rating results on a predetermined scale within the corresponding ranking classification; Y_k is a set of marked evaluation and ranking of user comments as the degree of permission to publish on the site/page, if necessary, with a prohibition mark for a specific contributor to write further comments and ranking by the degree of trust of all contributors. The list of the output flow of content, its main features, the corresponding classification, and IT generation/support/analysis contributes to the definition of precise general functional requirements for implementing the CLS to solve any NLP problem.

The model of the process of linguistic analysis of the Ukrainian-language text M_{lat} is presented

$$M_{lat} = \langle X, W, C, K, Y, D, S_{IAC}, S_{LA}, S_v, S_{\varpi_1}, S_{\varpi_2}, S_{\rho_1}, S_{\rho_2}, S_{\rho_3}, S_{\rho_4}, S_v, v, \varpi_1, \varpi_2, \rho_1, \rho_2, \rho_3, \rho_4, v \rangle,$$

where X is the input data in the CLS from various sources of information W ; Y is the original relevant content from the CLS as a result of the IIS according to the requests of users/visitors; S_{LA} is the process of linguistic analysis of content as a component of the IATCF subsystem S_{IAC} ; S_v is the process of generation/modification of the rules of operation of all modules by the moderator of the CLS; S_{ϖ_1} is the process of filling an unstructured database with integrated content X ; S_{ϖ_2} is the filling module of the structured database based on the processed integrated content C ; S_{ρ_1} and S_{ρ_2} are processes of generating results according to the requests of visitors and users; S_{ρ_3} is a cache processing process for generating reports on popular requests from CLS users; S_{ρ_4} is cache filling/modification process; S_v is the process of generating statistical results of the functioning of the CLS/modules and the activities of users D ; v is the operator of generation/modification of the rules of operation of all modules from the moderator of the CLS; ϖ_1 is the operator of filling an unstructured database with integrated content X ; ϖ_2 is the operator of filling the structured database based on the processed, integrated content of C ; ρ_1 and ρ_2 are operators for generating results according to the requests of visitors and users; ρ_3 is a cache processing operator for generating reports Y on popular requests from users; ρ_4 is cache filling/modification operator with K data; v is an operator for generating statistical results of the functioning of the CLS/modules and user activities:

$$S_{LA} = \langle X, Y, C, D, R, \alpha, \beta, \gamma, \delta, \lambda, o, \iota, \varsigma, \mu \rangle, \quad Y = \mu \circ o \circ \varsigma \circ \iota \circ \lambda \circ \delta \circ \gamma \circ \beta \circ \alpha, \quad (18)$$

where X is the input text data array; Y is a tuple of the original processed text according to the purpose of the CLS; C is a set of intermediate content, which is processed at the appropriate level in the CLS; D is auxiliary dictionaries; R is a set of processing rules; α is grapheme analysis operator (GA); β is morphological analysis operator (MA); γ is lexical analysis operator (LA); δ is operator of syntactic analysis (SA); λ is semantic analysis operator (SEM); o is ontological analysis operator; ι is reference analysis operator; ς is structural analysis operator; μ is operator pragmatic analysis (PA).

The primary process of linguistic analysis of textual content is presented:

$$Y = \mu(C_\mu, D_\mu, R_\mu, o(C_o, D_o, R_o, \varsigma(C_\varsigma, D_\varsigma, R_\varsigma, \iota(C_\iota, D_\iota, R_\iota, \lambda(C_\lambda, D_\lambda, R_\lambda, \delta(C_\delta, D_\delta, R_\delta, \gamma(C_\gamma, D_\gamma, R_\gamma, \beta(C_\beta, D_\beta, R_\beta, \alpha(C_\alpha, D_\alpha, R_\alpha, X))))))))), \quad (19)$$

where the content sets $C = \{C_\mu, C_o, C_\varsigma, C_\iota, C_\lambda, C_\delta, C_\gamma, C_\beta, C_\alpha\}$, linguistic dictionaries $D = \{D_\mu, D_o, D_\varsigma, D_\iota, D_\lambda, D_\delta, D_\gamma, D_\beta, D_\alpha\}$ and sets of production/association rules $R = \{R_\mu, R_o, R_\varsigma, R_\lambda, R_\delta, R_\gamma, R_\beta, R_\alpha\}$.

The primary linguistic process of processing textual Ukrainian-language information to solve a specific task of the NLP consists of nine stages:

Stage 1. Grapheme analysis α of textual Ukrainian-language information X :

$$C_\alpha = \alpha(X, D_\alpha, R_\alpha), \quad C_\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (20)$$

where X is the input text data array; α is GA operator; C_α is grapheme structure of the input text; D_α is grapheme dictionaries and libraries; R_α is GA rules; α_1 is an optical character recognition operator; α_2 is grapheme parsing operator of the input text X into sections, paragraphs and sentences; α_3 is grapheme analysis operator of linguistic chains into separate words; α_4 is the operator for forming a set of unrecognized chains; α_5 is the operator of identification and marking of unrecognized chains as numbers, dates, constant returns, abbreviations, proper and geographical names, etc.; α_6 is the operator for marking non-text strings as special symbols, formulas, figures, tables, etc.; α_7 is an operator for generating a marked linear sequence of words C_α with official signs and connections.

Stage 2. Morphological analysis β of text content C_β consists in the identification, analysis and determination of the form and structure of words, in particular:

$$C_\beta = \beta(C_\alpha, D_\beta, R_\beta), \quad C_\beta = \beta_3 \circ \beta_2 \circ \beta_1 \text{ or } C_\beta = \beta_3 \circ \beta_4 \circ \beta_1, \quad (21)$$

where β_1 is the morphological segmentation operator of the grapheme-recognized chain of symbols (words/tokens); β_2 is a token lemmatization operator; β_3 is the operator for marking parts of speech for segmented words; β_4 is the word stemming operator.

Production rules for identification/generation of Ukrainian participles [51]:

I. Formation of grammatical meanings: $\{D_K \rightarrow D_K(x, y)\}$, where $x = (act/pas)$; $y = (pres/past)$, for example, $\{D_K \rightarrow D_K(pas, pres)$; $D_K \rightarrow D_K(act, pres)$, ...}.

II. Analysis of morphemes: $\{D_K(act, pres) \rightarrow O'(t, \bar{d}, a_3)C(act, pres, a_3)\Phi$; $D_K(act, past) \rightarrow O'(\bar{t}, d, a_3)C(act, past, a_3)\Phi$; $D_K(pas, pres) \rightarrow O'(t, d - \bar{d}, a_3)C(pas, pres, a_3)\Phi$; $D_K(pas, past) \rightarrow O'(t, d - \bar{d}, a_3)C(pas, past, a_3)\Phi\}$, where O , C , Φ are designation of various morphemes without description.

III. Decomposition of the verb stem: $\{O'(\overline{atem}) \rightarrow O(\overline{atem})T$; $O'(\bar{d}, \bar{\emptyset})C(x, y) \rightarrow O(\bar{d}, \bar{\emptyset})C_d C(x, y, I)$; $O'(atem) \rightarrow O(atem)\}$, where T is thematic element (TE) $-и(i, \bar{i})/-а(я)/-ол(р)о-$; \overline{atem} is attribute value a_4 different from $atem$, i.e. $(a/i/\bar{a}/\bar{i}/o)$, C_d is verb suffix; $\bar{\emptyset}$ is any attribute value other than \emptyset ; x and y must satisfy the following condition: at $x = pas$ it is necessary that $y = pres$.

IV. TE identification: $\{(\bar{a})T\alpha \rightarrow O(\bar{a})\zeta$; $O(\bar{i})T\alpha \rightarrow O(\bar{i})\zeta$; $O(a)T \rightarrow O(a)a +$; $O(i)T \rightarrow O(i)i +$; $O(o)T \rightarrow O(o)o +$; $O(\bar{d}, II, a)TC(act, pres) \rightarrow O(\bar{d}, II, a)a + C(act, pres)$; $O(d - \bar{d}, I, a)TC(pas, pres) \rightarrow O(d - \bar{d}, I, a)a + C(pas, pres)$; $O(d - \bar{d}, I, i)TC(pas, pres) \rightarrow O(d - \bar{d}, I, i) + C(pas, pres)$; $(\bar{a}, II)T\beta \rightarrow O(\bar{a}, II)a + \xi$; $O(\bar{i}, I)T\beta \rightarrow O(\bar{i}, I) + \xi\}$, where ζ and ξ are arbitrary vowel and consonant; $+$ is boundary between morphemes.

V. Forming verbs with the appropriate morpheme: $\{O(I, y)C_d \rightarrow O(I, y)ува +$; $O(I, y)C_d \rightarrow O(I, y)ова +$; $O(\bar{y})C_d \rightarrow O(\bar{y})$; $O(\bar{t}, d, H)C_d \rightarrow O(\bar{t}, d, H) + C(pas, past)$; $O(t, d, H)C_d C(pas, pres) \rightarrow O(t, d, H)Hy + C(pas, pres)\}$.

VI. Suffix identification: $\{C(act, past, I - II) \rightarrow л +$; $O(atem)C(act, pres, I) \rightarrow уч +$; $O(\overline{atem})YC(act, pres, I) \rightarrow юч +$; $O(atem)C(act, pres, II) \rightarrow ач +$; $O(\overline{atem})YC(act, pres, II) \rightarrow яч +$; $C(pas, pres/past, I - II) \rightarrow H +$; $C(pas, pres/past, I - II) \rightarrow T +$; $O(atem)C(pas, pres/past, I - II) \rightarrow еH +$; $O(\overline{atem})YC(pas, pres/past, I - II) \rightarrow еH +$; $O(atem)C(pas, pres/past, I - II) \rightarrow ува +$; $O(\overline{atem})YC(pas, pres/past, I - II) \rightarrow юва +$; $C(pas, pres/past, I - II) \rightarrow овува +$; $O(atem)C(pas, pres/past, I - II) \rightarrow ова +$;

$O(\overline{atem})YC(pas, pres/past, I - II) \rightarrow$ йова +; $O(\overline{atem})X'C(pas, pres/past, I - II) \rightarrow$ X'ьова+}, where Y is any suffix/TE; X' is soft consonant, X is arbitrary consonant.

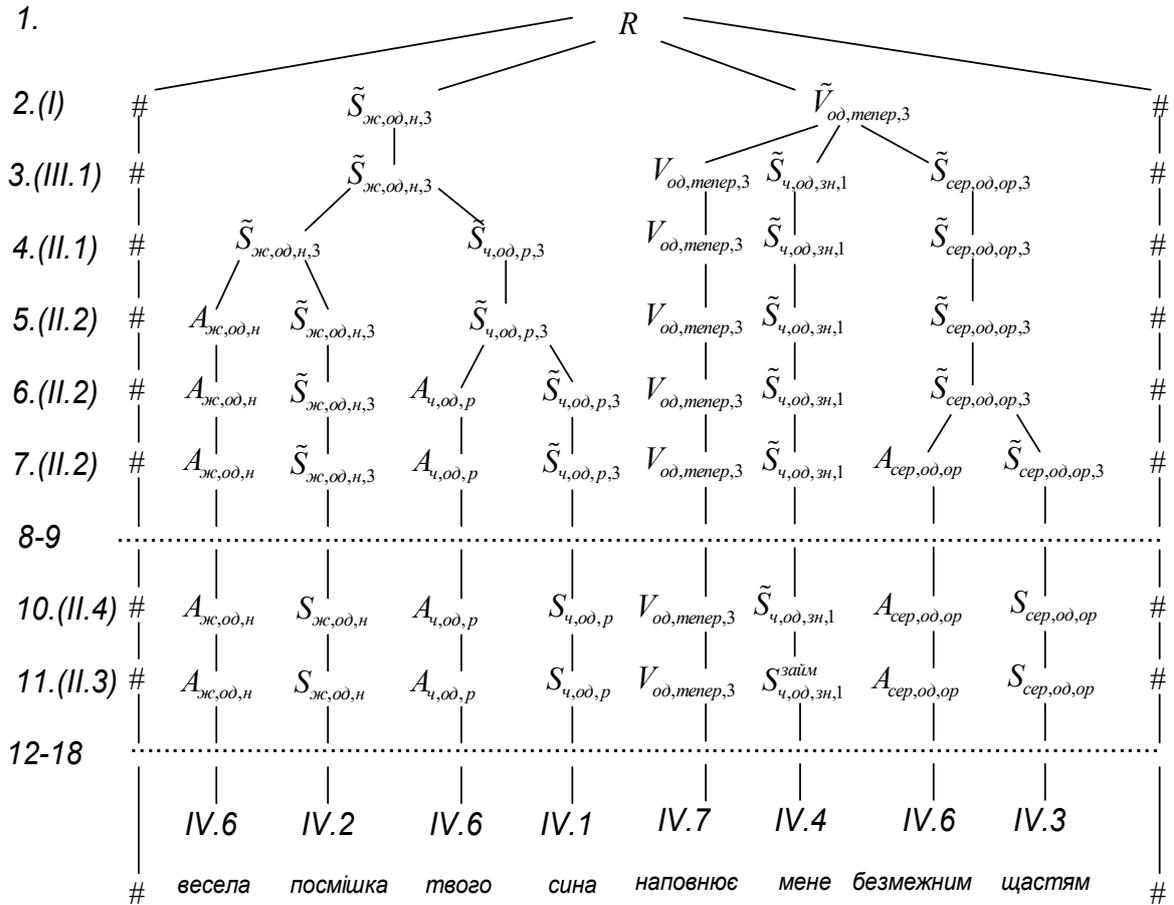


Figure 2: An example of building a tree for parsing the dependencies of sentence words

VII. Selection of the verb form (f/\bar{f}) and inflection: $\{\Phi \rightarrow \Phi(f); \Phi(f, s) \rightarrow \text{ого, им, ому}; \Phi(f, m) \rightarrow \text{ий}; \Phi(f, w) \rightarrow \text{ою, ої}; \Phi(f, \bar{s}) \rightarrow \text{им, ими, их}; \Phi \rightarrow \Phi(\bar{f}); \Phi(\bar{f}, w) \rightarrow \text{а, у}; \Phi(\bar{f}, k) \rightarrow \text{е}; \Phi(\bar{f}, \bar{s}) \rightarrow \text{і}; C(pas)\Phi(\bar{f}) \rightarrow \text{о}\}$.

VIII. Dictionary-based stem identification: $\{O(t - \bar{t}, d - \bar{d}, I, atem, y) \rightarrow \text{автоматиз+}, \text{буд+}, \text{мал'+}, \dots; O(t - \bar{t}, \bar{d}, I, atem, \emptyset) \rightarrow \text{вес+}, \dots; O(t, d - \bar{d}, II, \bar{i}, \emptyset) \rightarrow \text{втрач+}, \dots; O(\bar{t}, \bar{d}, I, a, \emptyset) \rightarrow \text{втруч+}, \dots; O(t, d - \bar{d}, I, \bar{i}, \bar{y}) \rightarrow \text{дослідж+}, \dots; O(\bar{t}, d, I, \bar{i}, \bar{y}) \rightarrow \text{запізн+}, \dots; O(t, \bar{d}, I, a, \emptyset) \rightarrow \text{кох+}, \dots; O(t, \bar{d}, II, \bar{i}, \emptyset) \rightarrow \text{люб+}, \dots; O(t, \bar{d}, I, atem, \emptyset) \rightarrow \text{нес+}, \dots; O(t, d, II, \bar{i}, \emptyset) \rightarrow \text{поділ+}, \dots; O(t, d, I, atem, \emptyset) \rightarrow \text{привес+}, \dots; O(t, d, I, atem, y) \rightarrow \text{побуд+}, \text{розфарб+}, \dots; O(\bar{t}, \bar{d}, I, \bar{a}, \emptyset) \rightarrow \text{сміj+}, \text{стогн+}, \dots; O(t, \bar{d}, I, a, \emptyset) \rightarrow \text{спит+}, \dots; O(\bar{t}, d, I, atem, н) \rightarrow \text{усміх+}, \dots; O(t, \bar{d}, I, atem, y) \rightarrow \text{фарб+}, \dots; O(t, d, I, о, \emptyset) \rightarrow \text{мол+}, \dots; O(\bar{t}, d, I, і, \emptyset) \rightarrow \text{змарн+}, \dots\}$.

IX. Basic morphological rules: $\{\alpha_1 + \rightarrow \alpha_1 + j\alpha_2; j + \text{и} \rightarrow \text{і}; \text{oZ} + C(pas, pres) + \Phi \rightarrow \text{aZ} + C(act, pres) + \Phi; \text{c}' + \text{W} \rightarrow \text{ш} + \text{W}; \text{в}' + \text{W} \rightarrow \text{вл}' + \text{W}; \text{б}' + \text{W} \rightarrow \text{бл}' + \text{W}; \text{д}' + \text{W} \rightarrow \text{дж}' + \text{W}; \text{т}' + \text{W} \rightarrow \text{ч} + \text{W}; \dots; \text{д} + \text{W} \rightarrow \text{д}' + \text{W}; \text{с} + \text{W} \rightarrow \text{с}' + \text{W}; \dots; \text{нн} + \Phi \rightarrow \text{н} + \text{o}\}$, where α_1 and α_2 are arbitrary vowels; j is sound designation [j] (йот); Z is any sequence not longer than 3 characters; W = -e(e)н-, -у(ю)ва-, -ова-, -овува-.

X. Graphical and orthographic rules: $\{j + a \rightarrow \text{я}, ja \rightarrow \text{я}; j + y \rightarrow \text{ю}, jy \rightarrow \text{ю}; j + e \rightarrow \text{є}, je \rightarrow \text{є}; \dots; X' + a \rightarrow X + \text{я}; X' + y \rightarrow X + \text{ю}; X' + и \rightarrow X + \text{и}; X' + i \rightarrow X + \text{і}; X' + e \rightarrow X + \text{є}\}.$

XI. Erasure of the boundary indicator between morphemes: $\{A + B \rightarrow AB\}$, where A and B are any morphemes that none of the rules of groups IX-X apply to $A + B$.

Stage 3. Lexical analysis γ of the text content C_γ in the intermediate stage of the analysis of the lexeme sequence to generate a parsing tree at the SA level:

$$C_\gamma = \gamma(C_\beta, D_\gamma, R_\gamma), \quad C'_\gamma = \gamma_2 \circ \gamma_1, \quad C''_\gamma = \gamma_5 \circ \gamma_4 \circ \gamma_3 \text{ or } C'_\gamma = \gamma_5 \circ \gamma_4, \quad (22)$$

where γ_1 is a speech segmentation operator for identification/clarification of words/phrases/tokens after MA; γ_2 is speech recognition or speech-to-text operator; γ_3 is optical character recognition operator as the second part after GA and MA for clarifying incorrect moments of recognition, taking into account the recognized adjacent tokens; γ_4 is the word tokenization/segmentation operator as data preparation for building a parsing tree at SA; γ_5 is text-to-speech.

Stage 4. The syntactic analysis δ of text content C_δ consists in building a tree for parsing word dependencies (Fig. 2) in a sequence of lexemes based on their categories:

$$C_\delta = \delta(C_\gamma, D_\delta, R_\delta), \quad C_\delta = \delta_3 \circ \delta_2 \circ \delta_1, \quad (23)$$

where δ_1 is grammar induction implementation operator; δ_2 is the operator of identification/elimination of boundary ambiguity or sentence violation; δ_3 is operator of syntactic parsing of phrases/sentences for building a SA tree. Rules for formulating Ukrainian phrases:

I. Choice of structure: $\{R \rightarrow \#\tilde{S}_{x,y,n,w} \tilde{V}_{y,тепер,w}\# \}$, where \tilde{V} is verb group, \tilde{S} is noun group, x is gender, y is singular/од, or plural/мно; z is the case, w is the person.

II. Noun group: $\{\tilde{V}_{x,y,z,3} \rightarrow \tilde{S}_{x,y,z,3} \tilde{S}'_{x',y',p,w}; \quad \tilde{S}_{x,y,z,3} \rightarrow A_{x,y,z} \tilde{S}_{x,y,z,3}; \quad K_1 \tilde{S}_{x,y,z,w} K_2 \rightarrow K_1 S_{x,y,z,w}^{займ} K_2, K_1 \neq A_{x,y,z}, K_2 \neq \tilde{S}'_{z'}; \tilde{S}_{x,y,z,3} \rightarrow S_{x,y,z}\}.$

III. Verb group: $\{\tilde{V}_{y,тепер,w} \rightarrow V_{y,тепер,w} \tilde{S}'_{x',y',zn,w'} \tilde{S}''_{x'',y'',op,w''}; \quad \tilde{V}_{y,тепер,w} \rightarrow V_{y,тепер,w} \tilde{S}'_{x',y',op,w'} \tilde{S}''_{x'',y'',zn,w''}; \quad \tilde{V}_{y,тепер,w} \rightarrow V_{y,тепер,w} \tilde{S}'_{x',y',zn,w'} ; \quad \tilde{V}_{y,тепер,w} \rightarrow V_{y,тепер,w} \tilde{S}'_{x',y',op,w'}\}.$

IV. Substitution of words: $\{S_{ч,y,z} \rightarrow \text{сина}_{y,z}, \dots; S_{ж,y,z} \rightarrow \text{посмішка}_{y,z}, \dots; S_{сер,y,z} \rightarrow \text{щастя}_{y,z}, \dots; S_{x,од,z,1}^{займ} \rightarrow \text{я}_z; \quad S_{x,од,z,2}^{займ} \rightarrow \text{ти}_z; \quad V_{y,тепер,w} \rightarrow \text{наповнити}_{y,тепер,w}, \dots; \quad A_{x,y,z} \rightarrow \text{веселий}_{x,y,z}, \text{безмежний}_{x,y,z}, \text{мій}_{x,y,z}, \text{твій}_{x,y,z}, \dots\}.$

Stage 5. Semantic analysis λ of the Ukrainian-language text C_λ consists of

$$C_\lambda = \lambda(C_\delta, D_\lambda, R_\lambda), \quad C_\lambda = \lambda_2 \circ \lambda_1, \quad (24)$$

where λ_1 is the identification operator of lexical semantics with the generation of a collection of values of each lexeme of the text; λ_2 is the relational semantics identification operator of the interdependencies of the content of the lexemes of the text.

Stage 6. Reference analysis ι identification of interphase units C_ι .

$$C_\iota = \iota(C_\lambda, D_\iota, R_\iota). \quad (25)$$

Reference analysis is often part of SEM. For Ukrainian texts, when analysing large corpora of texts, it is best to carry out as a separate stage (for example, for the analysis of the correspondence of a social group/community in social networks or other dialogues to identify logical, meaningful connections between the posts of different participants due to the subjectivity of everyone's speech.

Stage 7. Structural analysis ζ of the Ukrainian-language text C_ζ based on the degree of coincidence of lexical, terminological units of unity of text fragments. It is often part of SEM for short texts/messages or not used at all. For large corpora of texts as an additional stage of elimination of marked inaccuracy in SEM.

$$C_{\zeta} = \zeta(C_l, D_{\zeta}, R_{\zeta}) \text{ or } C_{\zeta} = \zeta(C_{\lambda}, D_{\zeta}, R_{\zeta}). \quad (26)$$

Stage 8. Ontological analysis of o text content C_o on the basis or part of the results of SEM and reference/structural analyses if necessary:

$$C_o = o(C_{\zeta}, D_o, R_o), C_o = o(C_l, D_o, R_o) \text{ or } C_o = o(C_{\lambda}, D_o, R_o). \quad (27)$$

Stage 9. Pragmatic analysis of μ text content C_{μ} is used to determine the text's structure by considering the context of sentences when forming paragraphs, sections, and dialogues. PA is an essential addition to SEM, reference, and structural analyses if it does not contribute to eliminating marked inaccuracy.

$$Y = \mu(C_{\mu}, D_{\mu}, R_{\mu}, C_{\lambda}, [C_o, C_{\zeta}, C_l]), Y = \mu_2 \circ \mu_1, \quad (28)$$

where μ_1 is a semantics identification operator outside individual sentences/phrases; μ_2 is the operator of text processing through higher-level NLP applications, for example, to simulate intelligent behaviour and an apparent understanding of natural language.

A general scheme/model of the pipeline of the CLS operation has been developed based on improved methods of processing information resources such as integration, maintenance and content management, as well as the development of improved methods of intellectual and linguistic analysis of text flow using machine learning technology (Fig. 3) [52-58]. Based on feedback from the user and output data of the ML model, the target audience interacts with the CLS, which contributes to the adaptation of the selected learning model. Five stages of relevant processes determine the basic architectural principles of building a typical CLS. The methods of monitoring, developing and managing content are interaction, formatting/filtering, NLP, ML and data accumulation in DS. Content and support processes feature analysis, deployment, prediction, interpretation, and content/result presentation. At the interaction stage, a set of rules for integrating content from multiple reliable sources at certain intervals is developed. Also, in parallel, a set of rules for checking the data entered by the user of the CLS was created as a preliminary stage for the formatting/filtering stage according to a collection of rules and content from the DS set in advance by the moderator. The next stage of NLP is an intermediate stage for ML and data accumulation. The ML stage is implemented through SQL queries and modules. The support process is more accessible to implement than the management stage, especially when analysing the results of the NLP, in which additional lexical resources and artefacts (dictionaries, translators, regular expressions, etc.) are created, which directly depend on the effectiveness of the CLS functioning (Fig. 4) [52-58].

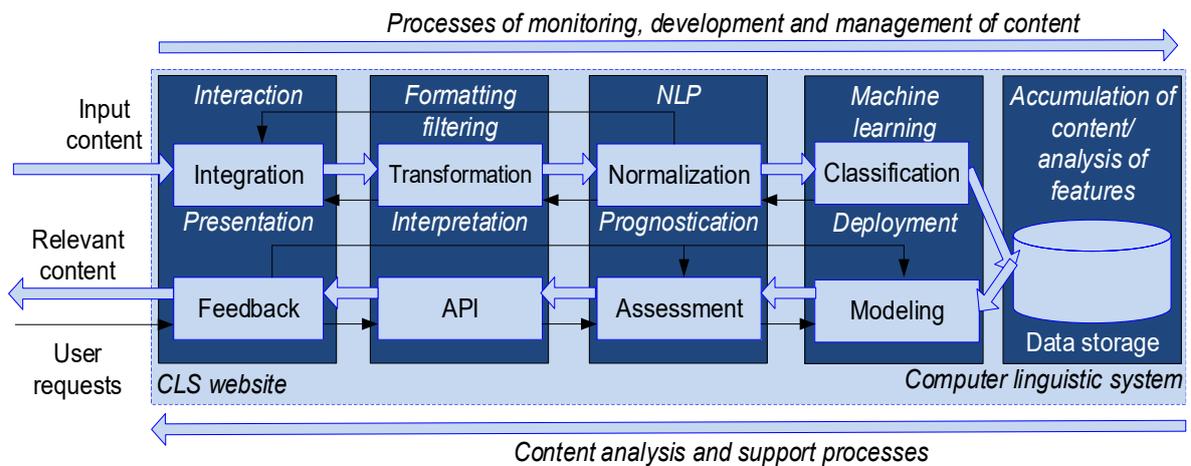


Figure 3: Scheme of the pipeline of the CLS operation

The transition process from the raw text to the expanded ML model consists of additional content transformations. First, the input text content is transformed into the input corpus as a collection of texts, accumulated and stored in the DS. The incoming content is further grouped, filtered, formatted,

linguistically processed, marked, normalized and converted into vectors for further processing. In the final transformation of the model (Fig. 5) [52-60], they train on the vector corpus to create a generalized presentation of the original content for further use in solving a specific NLP problem.

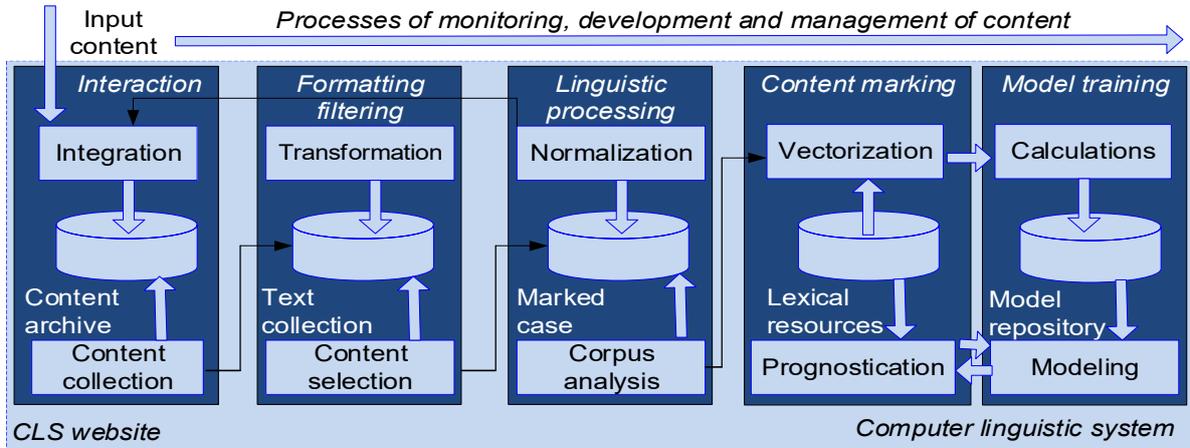


Figure 4: Scheme of the pipeline for processing Ukrainian-language textual content

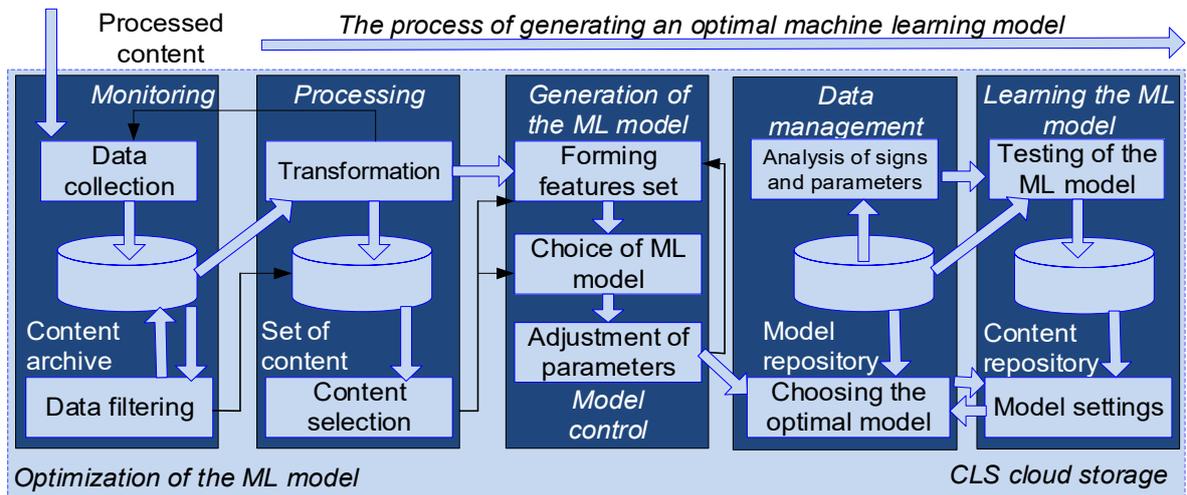


Figure 5: Machine learning pipeline process

NLP methods have been improved based on the developed 82 regular expressions (RGs) of pattern matching in GA and more than 2000 RGs of morphological analysis of Ukrainian-language texts. RV's primary admissible operations are the union and disjunction of symbols/chains/expressions, number and precedence operators, and anchors of the presence/absence of symbols in regular expressions. The main stages of tokenization and normalization of the Ukrainian text by cascades of simple substitutions of RG and finite automata are determined. Algorithms for word segmentation and normalization, sentence segmentation, and Porter's modified stemming are implemented and described as an effective way of identifying lem affixes for the possibility of marking the analysed word. Porter's modified stemming algorithm is based on searching/checking the obtained intermediate results with the tree of inflexions (so as not to go through all possible inflexions) and with the content of thematic dictionaries of bases with a set of PG-rules for identification of features (classification by parts of speech).

Step 1. Identify the next lexeme as the word w_i ($w_s = w_i$).

Step 2. Check with the stop word dictionary whether $D_{w_{sw}}$ or w_s is a service word. If yes, then $i = i + 1$ and go to step 1. Otherwise, go to step 3.

Stage 3. Go to the end of the word w_s . Recognize the inflection f_1^i in w_s from all possible ones (the longest one is chosen, for example, in w_s =текстова we choose the ending f_1^i =ова, not $f_1^i \neq a$)

from the RG of the word type $R_{adjectival}$, R_{noun} , or R_{verb} and in the presence of deletion of the inflexion f_1^i .

Stage 4. Saving the inflexion f_1^i in the word tag w_i .

Stage 5. Label w_i as type $m_{adjectival}^{w_i}$, $m_{noun}^{w_i}$ or $m_{verb}^{w_i}$, respectively.

Stage 6. Finding the deleted inflexion f_1^i in the tree of inflexions $T_{flection}$ (the longest one is chosen). Checking the contents of the subtree $T_{flection}^{f_1}$ with the existing word ending f_2^i ($f = f_2^i + f_1^i$). If w_s ends in f_2^i and has a counterpart in $T_{flection}^{f_1}$, then we store it in $f_i = f$ and delete in w_s .

Stage 7. We check the obtained base w_s of the initial word w_i with the content of the dictionary of bases D_{w_s} of words of the Ukrainian language. If there is no respondent, we store $\langle w_i, w_s \rangle$ in the additional temporary intermediate dictionary $D_{\langle w_i, w_s \rangle}$ for the moderator and proceed to stage 1. Otherwise, proceed to stage 4.

Stage 8. Analysis of inflexion and the presence/absence of alternation of letters in the base/inflexions of the words $\langle w_i, w_s \rangle$ and the analogue of the base of the word in D_{w_s} according to the corresponding PG-rule of MA to identify additional features of the analyzed word w_i .

Stage 9. Adding the identified linguistic features of the recognized part of speech to the tag of the word w_i of the type $m_{adjectival}^{w_i}$, $m_{noun}^{w_i}$ or $m_{verb}^{w_i}$, respectively. Saving the results in the corresponding dictionary D_{w_i} of the analysed text.

Unlike the classic Porter's algorithm, the modified one is adapted specifically for the Ukrainian language and gives an accurate result in 85-93% of cases, depending on the quality, style, genre of the text and, accordingly, the content of the dictionaries of CLS. In total, about 1,300 rules for processing suffixes and endings, considering the alternation of letters, adjectives - 99 RG-rules, and verbs - more than 800 RG-rules have been implemented for MA Ukrainian-language nouns. The algorithm for the minimum editorial distance of lines of Ukrainian texts is described as the minimum number of operations required to transform one into another. Also, an algorithm for calculating the maximum likelihood metric for the 2-gram and 3-gram models based on the analysis of word bases was developed to identify stable word combinations as keywords. To forecast the conditional probability of the following base of the word, we use the Markov assumption (the probability of the word depends on the previous one).

Moreover, suppose the keywords are a set of nouns or an adjective with a noun. In that case, other words, such as verbs, participles, etc., will be considered additional separators as other punctuation marks that demarcate persistent phrases as potential keywords. The order of bases is not crucial for the Ukrainian language.

Stage 1. Process the input text and break it into separate phrases (sentences) $R_1 R_2 \dots R_m$, marking each start-end with the corresponding $\langle p \rangle \langle /p \rangle$ tag. Eliminate all non-alphabetic characters. Convert uppercase letters to lowercase. Remove official words if necessary (for certain NLP tasks).

Stage 2. Apply Porter's stemming to obtain the sequence of word stems $x_{i1} x_{i2} \dots x_{in_i}$ of word stems $\forall R_i$ taking into account word normalization, respectively.

Stage 3. Receive input queries $Q_1 Q_2 \dots Q_k$ as a sequence of words of the searched data. Find $\forall Q_j$ for each word $y_{j1} y_{j2} \dots y_{jk_j}$ basis by stemming.

For example, for the search phrase Q_j :

*Методи та засоби опрацювання інформаційних ресурсів
систем електронної контент комерції*

Translation - Method and tools for information systems processing in electronic content commerce systems

y_{j1}	y_{j2}	y_{j3}	y_{j4}	y_{j5}	y_{j6}	y_{j7}	y_{j8}	y_{j9}	y_{j10}
метод	та	засіб	опрац	інформ	ресурс	систем	електрон	контент	комерц
58	190	25	62	122	83	170	89	408	300

Stage 4. Conduct a statistical analysis of the occurrence of word stems and sequences of query word stems in the analyzed text.

The text	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	x_{i9}	x_{i10}
Words basics	метод	та	засіб	опрац	інформ	ресурс	систем	електрон	контент	комерц
x_{i1}	метод	0	8	0	6	0	0	0	1	0
x_{i2}	та	2	0	5	1	7	0	2	0	1
x_{i3}	засіб	0	2	0	14	0	0	0	0	0
x_{i4}	опрац	0	0	0	0	46	0	0	1	3
x_{i5}	інформ	0	0	0	0	0	64	9	0	0
x_{i6}	ресурс	0	7	0	0	0	0	0	1	0
x_{i7}	систем	0	8	0	1	0	0	0	21	0
x_{i8}	електрон	0	0	0	0	0	0	0	0	72
x_{i9}	контент	0	10	0	0	0	0	0	0	73
x_{i10}	комерц	0	6	0	0	0	0	0	0	176

Stage 5. Find the probability of the appearance of 2-grams in the analyzed text. In each row, the value is divided by y_{ji} , where i is the row number after normalization.

The text	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	x_{i9}	x_{i10}	y_{ji}
words basics	метод	та	засіб	опрац	інформ	ресурс	систем	електрон	контент	комерц	
x_{i1}	метод	0	0.18	0	0.1	0	0	0	0.02	0	58
x_{i2}	та	0.01	0	0.03	0.005	0.035	0	0.01	0	0.005	190
x_{i3}	засіб	0	0.08	0	0.16	0	0	0	0	0	25
x_{i4}	опрац	0	0	0	0	0.74	0	0.016	0.048	0.064	62
x_{i5}	інформ	0	0	0	0	0	0.52	0.074	0	0	122
x_{i6}	ресурс	0	0.08	0	0	0	0	0.012	0	0	83
x_{i7}	систем	0	0.05	0	0.006	0	0	0.124	0	0	170
x_{i8}	електрон	0	0	0	0	0	0	0	0.81	0.112	89
x_{i9}	контент	0	0.03	0	0	0	0	0	0	0.179	408
x_{i10}	комерц	0	0.02	0	0	0	0	0	0.053	0	300

The resulting matrices will, in most cases, be sparse. Phrase and various variations (plural/singular and cases) $P(\text{система електронної контент комерції})$:

$$P(\text{електрон}|\text{систем})P(\text{контент}|\text{електрон})P(\text{комерц}|\text{контент}) = 0.124 \times 0.81 \times 0.179 = 0.01797876.$$

The SEM method has been improved based on the taxonomy of concepts, which specifies the syntax of the Ukrainian language as the root concept of the ontology: $Concepts_{\mu}: < R_{Snt} > \rightarrow C'_{\mu}$.

In SEM, to identify the set of semes of the corresponding Ukrainian-language text and their relationship, first, based on the results of SA, a semantic graph of the relations of linguistic units is built, taking into account the parts of the language of words:

$$C'_{\mu} = \lambda(C_{\lambda}, D_{\lambda}, R_{\lambda}, Concepts_{\mu}), Concepts_{\mu} = \langle C_{WrdCmb}, C_{SntCmb} \rangle,$$

where C_{WrdCmb} is a tuple of concepts of phrase formation; C_{SntCmb} is a tuple of sentence generation concepts in the Ukrainian language. Tuple C_{WrdCmb} is given as:

$$C_{WrdCmb} = \langle Sgn_1^{Wrd}, Sgn_2^{Wrd}, Sgn_3^{Wrd}, Sgn_4^{Wrd} \rangle,$$

where Sgn_i^{Wrd} is a tuple of phrase generation properties:

$$Sgn_1^{Wrd} = \langle Sgn_{Lxc}^I, Sgn_{Snt}^I \rangle, \\ Sgn_2^{Wrd} = \langle Sgn_{Now}^{II}, Sgn_{Adc}^{II}, Sgn_{Nmr}^{II}, Sgn_{Prn}^{II}, Sgn_{Vrb}^{II}, Sgn_{Adv}^{II} \rangle, \\ Sgn_3^{Wrd} = \langle Sgn_{Crd}^{III}, Sgn_{Inf}^{III} \rangle, Sgn_4^{Wrd} = \langle Sgn_{SmWd}^{IV}, Sgn_{CmWd}^{IV} \rangle,$$

where Sgn_{Lxc}^I is a tuple of lexical signs of phrase generation; Sgn_{Snt}^I is a tuple of syntactic signs of phrase generation; Sgn_{Nou}^{II} is a tuple of named properties; Sgn_{Adv}^{II} is a tuple of adjectival properties; Sgn_{Nmr}^{II} is a tuple of properties of numerals; Sgn_{Prn}^{II} is a tuple of pronominal properties; Sgn_{Vrb}^{II} is a tuple of verb properties; Sgn_{Adv}^{III} is a tuple of adverbial properties; Sgn_{Crd}^{III} is a tuple of consecutive properties and Sgn_{Inf}^{III} is a tuple of subordinate properties; Sgn_{SmWd}^{IV} is a tuple of ordinal properties and Sgn_{CmWd}^{IV} is a tuple of subordinate properties.

The tuple Sgn_{Crd}^{III} describes the component properties of a relation clause:

$$Sgn_{Crd}^{III} = \langle Sgn_{AdCm}^{Crd}, Sgn_{CnCm}^{Crd}, Sgn_{DvCm}^{Crd} \rangle,$$

where Sgn_{AdCm}^{Crd} is a tuple of the properties of a separating connection, Sgn_{CnCm}^{Crd} is a tuple of the properties of a connecting connection, and Sgn_{DvCm}^{Crd} is a tuple of the properties of an opposing connection.

$$Sgn_{Inf}^{III} = \langle Sgn_{CtCm}^{Inf}, Sgn_{MgCm}^{Inf}, Sgn_{AgCm}^{Inf} \rangle,$$

where Sgn_{CtCm}^{Inf} is a tuple of matching properties; Sgn_{MgCm}^{Inf} is a tuple of control properties; Sgn_{AgCm}^{Inf} is a tuple of adjacency properties. A tuple of sentence generation concepts: $C_{SntCmb} = \langle Sgn_1^{Snt}, Sgn_2^{Snt}, Sgn_3^{Snt}, Sgn_{SntMb}^{Snt} \rangle$, where sentence generation properties are grouped in Sgn_i^{Snt} are a tuple of sentence generation properties; Sgn_{SntMb}^{Snt} is a tuple of clause identification properties;

$$Sgn_1^{Snt} = \langle Sgn_{NrSn}^I, Sgn_{PrSn}^I, Sgn_{InSn}^I \rangle, Sgn_2^{Snt} = \langle Sgn_{EmNt}^{II}, Sgn_{EmCl}^{II} \rangle, \\ Sgn_3^{Snt} = \langle Sgn_{SlSt}^{III}, Sgn_{ClSt}^{III} \rangle, Sgn_{SntMb}^{Snt} = \langle Sgn_{MnStMb}^{SntMb}, Sgn_{SdStMb}^{SntMb} \rangle,$$

where Sgn_{NrSn}^I is a tuple of narrative sentence generation properties; Sgn_{PrSn}^I is a tuple of properties for generating interrogative sentences; Sgn_{InSn}^I is a tuple of prompt sentence generation properties; Sgn_{EmNt}^{II} is a tuple of properties for generating emotionally neutral sentences; Sgn_{EmCl}^{II} is a tuple of properties for generating emotional sentences; a tuple of concepts for the formation of Sgn_{SlSt}^{III} simple and Sgn_{ClSt}^{III} complex sentences; Sgn_{MnStMb}^{SntMb} is a tuple of properties identifying the main members of the sentence; Sgn_{SdStMb}^{SntMb} is a tuple of the properties of the identification of the secondary members of the sentence; $Sgn_{NrSn}^I = \langle Sgn_{AfSt}^{NrSn}, Sgn_{NgSt}^{NrSn} \rangle$; Sgn_{AfSt}^{NrSn} is a tuple of properties for generating affirmative sentences; Sgn_{NgSt}^{NrSn} is a tuple of negative sentence generation properties. To generate a simple sentence Sgn_{SlSt}^{III} features are analyzed:

$$Sgn_{SlSt}^{III} = \langle Sgn_1^{SlSt}, Sgn_2^{SlSt}, Sgn_3^{SlSt}, Sgn_4^{SlSt}, Sgn_5^{SlSt}, Sgn_6^{SlSt}, Sgn_7^{SlSt}, Sgn_8^{SlSt} \rangle,$$

where Sgn_i^{SlSt} is a tuple of simple sentence generation properties.

4. Experiments, results and discussion

I will analyse the results of the experimental approbation of the developed methods and means of linguistic, intellectual analysis of texts in the Ukrainian language based on the development of methods for identifying keywords, determining persistent word combinations, thematic classification of the text and detecting duplication of text. Let us consider the peculiarities of the process of syntactic analysis of Ukrainian-language textual content aimed at identifying significant keywords of input texts. Having determined the role and formal features of the syntactic analyser in the process of identifying keywords of the content topic, the procedures of the proposed method were decomposed into two stages (Table 1), where A (total keywords identified with a given word weight), B (generated significant words without pronoun and verbs), C (coincidence of words with the author's list), D (accuracy of the coincidence of identified keywords with the author's list), E (additionally defined keywords, but not determined by the author of the publication). In stage 1, the research for step 1 (analysis of full articles) and step 2 (articles without metadata such as abstract,

author keywords and list of references) was carried out without the application of ML, and in stage 2 - with ML. The method of article analysis without metadata achieves the best results according to the density criterion. The author of the article often defines a more significant number of words (A_2) and a smaller number of keywords (A_1) than are present in the text of the scientific and technical publication (Fig. 6). Unlike known parsers, the proposed method provides self-improvement and self-learning of the keyword definition module due to the identification mechanism of significant statistical parameters within the limits defined by the moderator. A system has been developed on the Victana website, which allows users to choose from a list of languages of the analysed text (<http://victana.lviv.ua/index.php/kliuchovi-slova>). The value of A_3 differs from the value of A_1 by 0.69 (by number, but not by content); A_4 from A_1 by 1.74; A_5 from A_1 by 2.66; A_6 from A_1 by 3.58. The value of A_2 differs from the value of A_3 by 4.36; respectively, A_2 from A_4 by 3.31; A_2 from A_5 by 2.39; A_2 from A_6 by 1.47. Adaptively changing the parameters/rules of the module almost doubles the collection of identified keywords (for example, the value of A_1 is greater than A_3 by 1.144654; A_6 by 1.750524; A_5 by 1.557652; A_4 by 1.36478). The total increase in value obtained depending on the moderation of dictionaries is, respectively, for A_3 is 14.46541; A_4 is 36.47799; A_5 is 55.7652; A_6 is 75.05241. When comparing A_2 is greater than $A_3 \div A_6$ and we have a chain of such values as 1.7985; 1.5084; 1.3217; and 1.176.

Table 1

Statistical data of the study of the content of scientific and technical publications

Name	Words weight	Stage 1					Stage 2				
		A	B	C	D	E	A	B	C	D	E
Step 1	≥ 1	5.46	3.92	2.51	2.08	1.74	7.43	7.03	3.27	3	4.18
	≥ 2	1.08	0.88	0.63	0.59	0.26	2.67	2.64	1.65	1.54	1.12
	≥ 3	0.41	0.38	0.22	0.21	0.16	1.21	1.2	0.85	0.79	0.41
	≥ 4	0.15	0.13	0.09	0.09	0.04	0.46	0.45	0.33	0.31	0.15
	≥ 5	0	0	0	0	0	0	0	0	0	0
Step 2	≥ 1	6.51	5.02	2.68	2.23	2.37	8.35	7.78	3.25	2.91	4.99
	≥ 2	1.34	1.11	0.74	0.72	0.39	3.12	3.07	1.81	1.67	1.43
	≥ 3	0.51	0.45	0.29	0.27	0.17	1.42	1.4	0.93	0.85	0.54
	≥ 4	0.19	0.17	0.12	0.12	0.05	0.73	0.72	0.45	0.42	0.31
	≥ 5	0.11	0.1	0.06	0.06	0.04	0.33	0.32	0.25	0.23	0.1

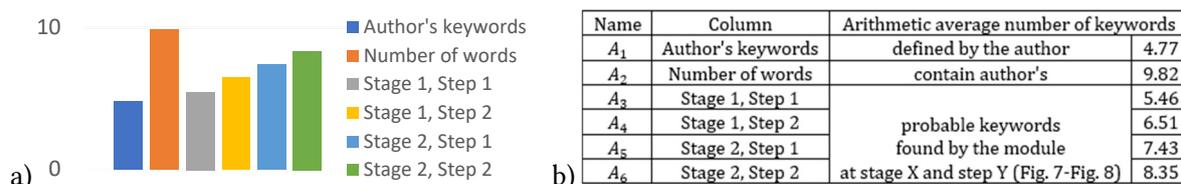


Figure 6: Results of the analysis of more than 300 scientific and technical publications

For different stages and steps of the experiment of processing the primary text, the average coincidence of the lists of discovered keywords with the author's keywords varies in the range of 52.6-68.5%. The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system ranges from 38.9-75.8%, depending on the stages of analysis of article texts. The accuracy of matching keywords compared to all found by the system varies between 34.3-71.9%, depending on the stages of analysis of article texts. For A_3 , the module most often identified the number of keywords {5, 7, 3} (≥ 10), although the distribution of found keywords was within [1;18] words (except 17).

For A_4 , the module most often identified the number of keywords also {5, 7, 3}, although the distribution of found keywords is within [1;18] (except 17), the number of identified words increased,

and the highest reliability index was achieved. For A_5 , the module most often identified the number of keywords {7, 6, 5, 10, 8}, although the distribution of found keywords was within [2;14] (the range narrowed significantly). For A_6 , the module most often identified the number of keywords {8, 5, 7, 10}, the distribution of identified keywords within [3;16] (accuracy improved). The accuracy of the definition of keywords increases during the moderation of dictionaries and the ML module. The difference between the number of keywords defined by the author and identified by the module at A_3 is 44.39919% (difference in %). Accuracy improves with A_4 is 33.70672%, significantly improving with A_5 is 24.33809%, and with A_6 is 14.96945%.

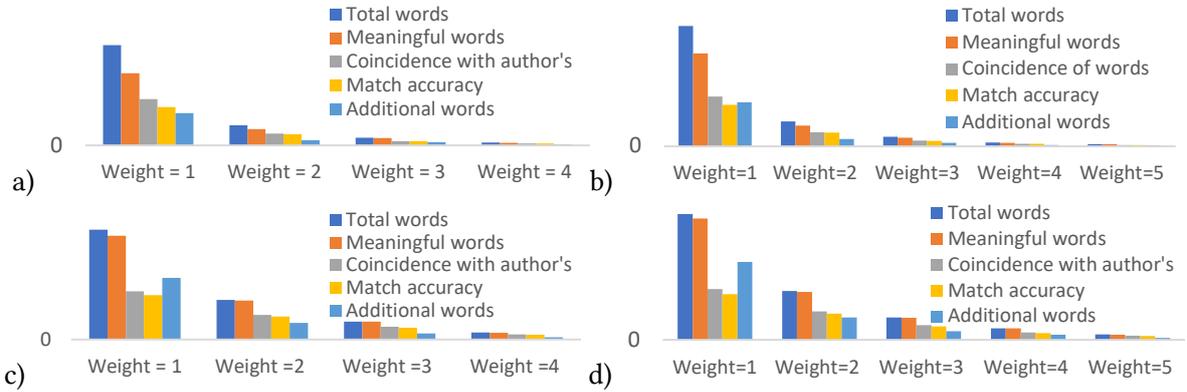


Figure 7: Obtaining meaningful words at the stage: a) 1.1, b) 1.2, c) 2.1 and d) 2.2

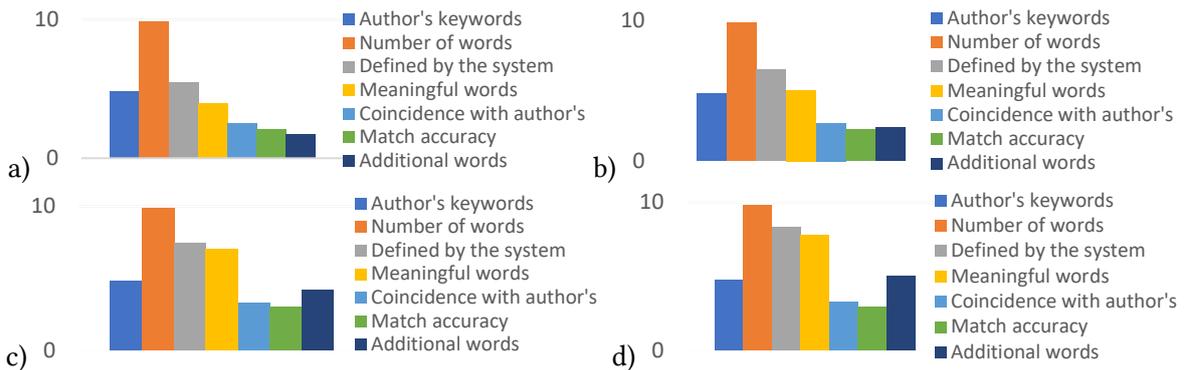


Figure 8: Arithmetic mean occurrence of words at the stage: a) 1.1, b) 1.2, c) 2.1 and d) 2.2

Analysis was performed for filtered texts without metadata and unfiltered texts. The average values obtained for filtered texts $\overline{Per}_f = 0.28$ and unfiltered $\overline{Per}_0 = 0.19$ shows that filtering scientific articles improves keyword density by 1.48 times or 47.83% (Fig. 9a).

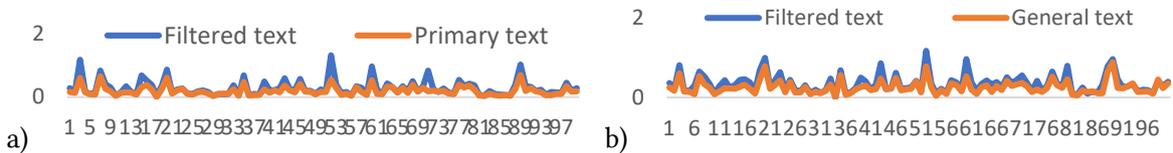


Figure 9: Results of checking articles without specifying the thematic dictionary

The obtained values for the texts $\overline{Per}_f^v = 0.34$ and $\overline{Per}_0^v = 0.25$, taking into account the refinement of the thematic dictionary through ML and the replenishment of blocked words, shows that filtering with simultaneous moderation of the thematic dictionary improves keyword density by 1.35 times or by 35.44% (Fig. 9b). A comparison of the values in the original author's text $\overline{Per}_0 = 0.19$ and $\overline{Per}_0^v = 0.25$ without/with the refinement of the thematic dictionary, respectively,

demonstrates the effectiveness of the moderation of the thematic dictionary in the initial text - the density of keywords increases 1.34 times or by 34.33% (Fig. 10a). Comparison of the values in the filtered author's text $\overline{Per}_f = 0.28$ and $\overline{Per}_f^p = 0.34$ without/with the refinement of the thematic dictionary, respectively, demonstrates the effectiveness of the moderation of the thematic dictionary in the filtered text as the density of keywords increases 1.23 times or by 23.14% (Fig. 10b).

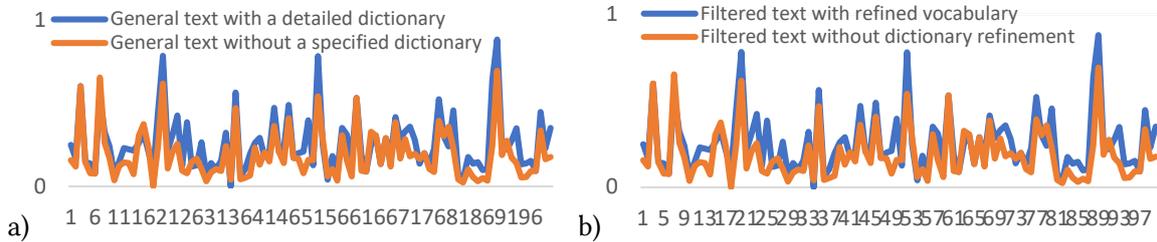


Figure 10: Results of analysis of articles with different dictionaries

So, the experimental study confirmed the method's reliability - for different stages of processing the primary text, the average coincidence of the lists of identified keywords with the author's keywords varies in the range of 52.6-68.5% (by 9%). The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system ranges from 38.9-75.8%, depending on the stages of analysis of article texts. The accuracy of matching keywords compared to all found by the system varies between 34.3-71.9%, depending on the stages of analysis of article texts. A method of determining stable word combinations when identifying textual content keywords in reference passages of the author's text has been developed. The process consists of the use of Zipf's law in the formation of stable word combinations as key, taking into account the following rules of preliminary linguistic processing of the text: removal of all stop words; form bigrams only within the limits of punctuation marks and words that are not verbs or pronouns (the latter are considered punctuation marks); determine verbs by inflexions; form bigrams based on their bases without taking into account their inflexions; definition of adjectives by inflexions and to believe that adjectives should only be in the first place in the bigram from Ukrainian-language texts. A module has been developed to identify persistent phrases as keywords in textual content. An approach to developing linguistic content analysis software for the determination of stable word combinations in identifying keywords of Ukrainian-language and English-language textual content is proposed. The peculiarity of the approach is adapting the linguistic, statistical analysis of lexical units to the peculiarities of the constructions of Ukrainian and English words/texts. The results of the experimental approbation of the proposed method of content analysis of English- and Ukrainian-language texts to determine stable word combinations in identifying keywords of technical texts were studied.

A method of identifying the style of the author of the text based on the analysis of linguistic speech coefficients in the standard has been developed. The technique consists of a comparative study of the author's attribution in the author's statistically processed work (standard) with an arbitrarily analysed passage. The method evaluates the probability of the text of the article belonging to the author of the benchmark with the analysis of the relevant coefficients of lexical speech as the concentration of the text I_{kt} , the coherence of the speech K_z , the uniqueness of the text I_{wt} , the syntactic complexity of the speech K_s and the linguistic diversity of the speech K_l . The degree of speech connectivity K_z does not decrease significantly. In 2001, it changed within [0.5; 1.2], and in 2021 – within [0.4; 0.9] (Fig. 11). Moreover, the method works under the condition that the author's standard has already been researched - the task of NLP is to form the author's frequency dictionary, including service/stop words.

An algorithm for determining stop words of text content based on linguistic analysis of text content has been developed. For the individual style of the author's text, markers are service/stop words (for example, particles, conjunctions, prepositions, parasite words, slang, slang, etc.) unrelated

to the article's topic. The absolute and relative frequencies of stopwords were analysed and compared with the reference values for each excerpt. Therefore, applying the method of reference words gives the following results: finding what most likely belongs to the standard among the studied passages. Other results also confirm the effectiveness of the keyword method in author attribution of texts. The proposed assumption about the insignificance of the influence of the share as a parameter of the process on the results led to a decrease in the correlation coefficients but placed the probability of belonging to the standard for passages in the correct order (Table 2). More likely, Excerpt 4 belongs to the author of the template (although there is no significant difference between results 4 and 2, if they are written in the same period, they do not belong to the author of the template; if in different periods with the template, the probability of belonging to this author increases).

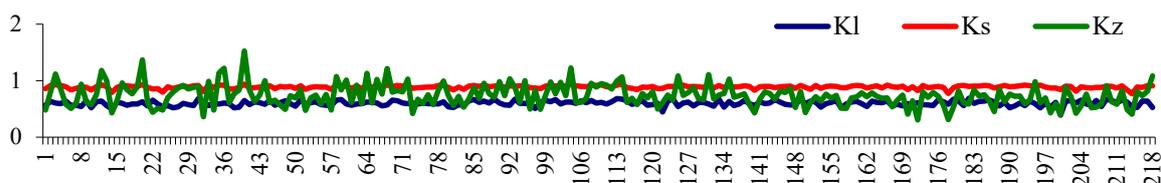


Figure 11: Analysis of the distribution of speech style parameters K_l , K_s and K_z

Table 2

Correlation coefficients for stop words

New numbering	Article number	R_{e-U}	Participle	Conjunction	Preposition	R'_{e-U}
1	4	0.7326	0.9594	0.9544	0.5639	0.6905
2	2	0.7066	0.9580	0.5714	0.4928	0.4913
3	1	0.6076	1	0.79	0.72	0.6900
4	3	0.2810	0.8800	0.1624	0.1517	0.2254

An algorithm for the linguistic analysis of Ukrainian-language texts and a syntactic analyser of text content has been developed. The features of the algorithm are the adaptation of morphological and syntactic analysis of lexical units to the peculiarities of constructions of Ukrainian words/texts. Algorithms are tested to identify significant stopwords in Ukrainian-language text based on regular expressions. When parsing words belonging to a part of speech, declension within this part of speech was taken into account. For this purpose, word inflexions were analysed for classification, selection of the basis and formation of the corresponding alphabetic-frequency dictionaries. The dictionaries contents were subsequently taken into account in the next steps of determining the text's authorship by calculating the parameters and coefficients of the author's speech. Software implementation for solving some NLP problems, as research of:

- keywords (<https://victana.lviv.ua/kliuchovi-slova>);
- stable phrases (<https://victana.lviv.ua/nlp/stiiki-slovoopoluchennia>);
- classification of textual content (<https://victana.lviv.ua/kliuchovi-slova>);
- quantitative evaluations of speech (<https://victana.lviv.ua/nlp/linhvometriia>);
- the author's style based on calculations of stylometry coefficients and their comparison with the corresponding coefficients in the standard text (<https://victana.lviv.ua/nlp/stylemetriia>);
- differences in text signs (<https://victana.lviv.ua/nlp/hlotokhronolohiia>);
- features of the style of texts based on N-grams (<https://victana.lviv.ua/nlp/n-grams>).

The results of the experimental approbation of the proposed content monitoring method for determining the author in Ukrainian-language scientific texts of a technical profile were studied. A comparison of the results of more than 300 one-person works of a technical direction by 100 different authors for 2001–2021 was carried out to determine whether and how the coefficients of text

diversity of these authors change in different periods. A method of identifying the potential (probable) author of a Ukrainian-language text based on the analysis of the author's linguistic speech coefficients in a reference passage of the author's text has been developed. Decomposition of the method of determining the author was carried out based on the analysis of such speech coefficients as speech coherence, degree of syntactic complexity, linguistic diversity, indices of concentration and exclusivity of the text. In parallel, such parameters of the author's style as the number of words in a specific text, the total number of words in this text, the number of sentences, the number of prepositions, the number of conjunctions, the number of words with a frequency of 1 and the number of words with a frequency of 10 and more, as well as keywords and 3 - grams. For example, 3-grams of 3 articles were analysed [61-63] (Ukrainian versions). For the most frequently used letters, the frequency of appearance of 3-grams with such initial letters will have an almost identical distribution (peak values in Fig. 12a), but not for other letters. Therefore, it is expedient to study only 3 grams for initial letters that occur less often in the texts of a specific language to determine the degree of belonging of the text to the corresponding author (for example, Fig. 12b). According to these graphs. It appears that Articles (1,2) are more likely to be written by the same author, although the same author could also write Articles (1,3) (but this is not true). Different authors write articles (2,3). Applying linguistic, statistical analysis of 3-grams to a set of articles makes it possible to form a subset of publications similar in terms of linguistic characteristics. Imposing additional conditions in the form of linguistic, statistical analyses (a set of keywords, stable word combinations (Table 3), stylometric, ligvometric, etc.) will significantly reduce the subset, clarifying the list of more likely authors' works. Thus, the analysis of the content and frequency of appearance of only official words separates Articles (1,3) into different subsets, leaving Articles (1,2) in one. 78.4814% of 3-grams were analysed for Article 1, 72.6332% for Article 2, and 84.1271% for Article 3. The difference in the use of the corresponding 3-grams between Articles (1,2) is $R_{12}=56.5254\%$, between Articles (2,3) – $R_{23}=69.4271\%$, between Articles (1,3) – $R_{13}=62.9839\%$. Accordingly, Articles (1,2) are more similar by [6-12]% ($R_{23}>R_{12}$ by 12.9017%, $R_{23} > R_{13}$ by 6.4432%, $R_{13}> R_{12}$ by 6.4585%, i.e. $R_{23}>R_{13}>R_{12}$) than Articles (1,3) and (2,3). The smaller the R_{ij} , the greater the degree to which the same author writes the articles. Then, in case Articles (1,2) are more likely to be written by one author/team than Articles (2,3) and (1,3), respectively.

Table 3

List by frequency rating of stable phrases for Article 1

Phrase	FREG		t-test		LR		X2	
	AF	RF	Phrase	t	Phrase	logL	Phrase	X2
система	4	0.08888	система	1.82222	інформаційний	5.03e	прийняття	45.00000
електронний		9	електронний	2	технологія	-1	рішення	0
інформаційни	4	0.08888	електронний	1.57809	інтелектуальни	2.13e	система	45.00000
й система		9	контент-	1	й система	-1	електронний	0
			комерція					
електронний	3	0.06666	розділ науковий	1.31993	інформаційний	8.36e	електронний	32.94642
контент-		7		3	система	-2	контент-	9
комерція							комерція	
розділ	2	0.04444	інформаційний	1.22222	портал	5.58e	розділ науковий	29.30232
науковий		4	система	2	науковий	-2		6
портал	1	0.02222	прийняття	0.97777	курс технологія	3.31e	курс технологія	21.98863
науковий		2	рішення	8		-2		6
інтелектуальн	1	0.02222	курс технологія	0.95555	сховище дані	3.31e	сховище дані	21.98863
ий система		2		6		-2		6
прийняття	1	0.02222	сховище дані	0.95555	прийняття	8.27e	портал	14.31818
рішення		2		6	рішення	-3	науковий	2
курс	1	0.02222	портал	0.93333	розділ науковий	1.89e	інформаційний	5.848550
технологія		2	науковий	3		-3	система	
сховище дані	1	0.02222	інтелектуальни	0.77777	електронний	1.55e	інтелектуальни	3.579545
		2	й система	8	контент-	-4	й система	
					комерція			
інформаційни	1	0.02222	інформаційний	0.68888	система	1.37e	інформаційний	1.890409
й технологія		2	технологія	9	електронний	-6	технологія	

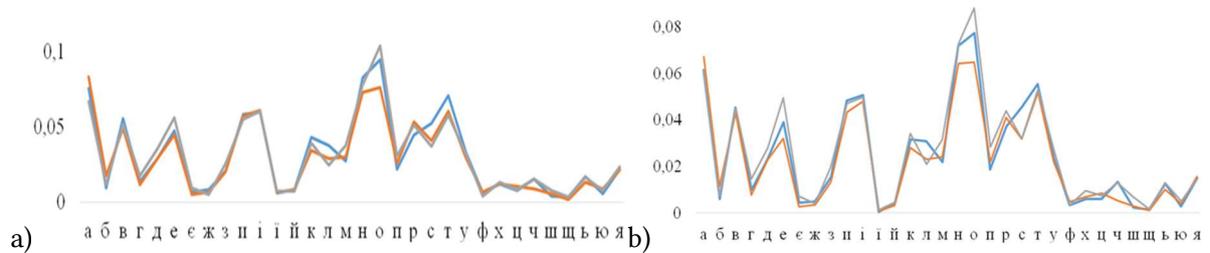


Figure 12: Graph of the frequency distribution of 1-gram and 3-gram occurrences in Articles 1–3 (blue for Article 1 [61], orange for Article 2 [62] and grey for Article 3 [63])

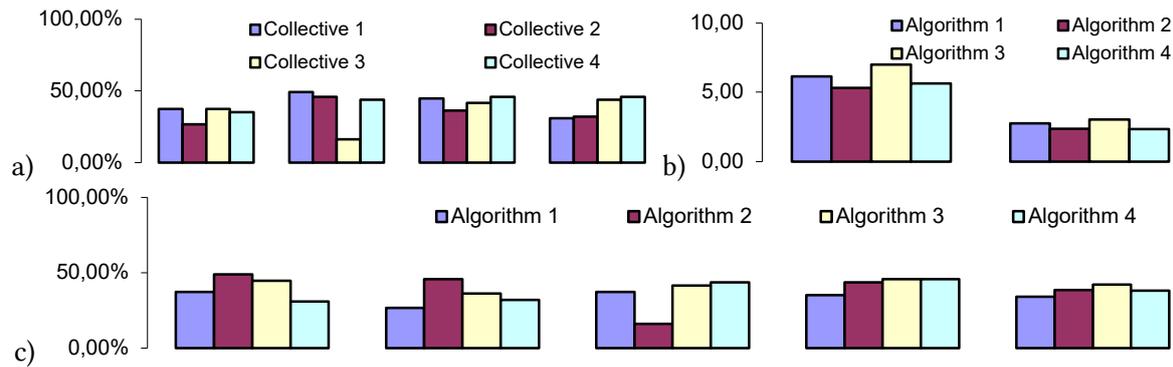


Figure 13: Style analysis: a – according to the developed algorithms 1-4; b – taking into account the signs of speech; c – for analysed collective works 1-4 and the average value

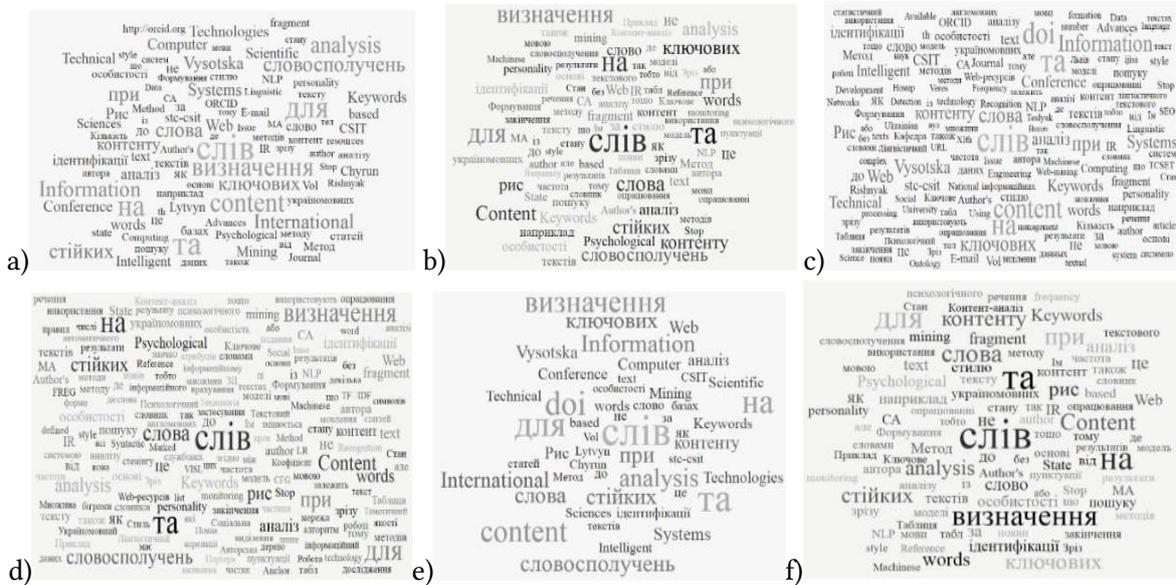


Figure 14: Study of style at stage 2 for the text with the construction of a frequency dictionary: a – complete with 100 words; b – the main one of 100 words; c – complete with 200 words; d – the main one of 200 words; e - complete with 50 words; f - the main one of 50 words

When identifying the author of a text, it is assumed that the text reflects the author's style of writing, which makes it possible to distinguish him from others. To compare texts with each other, it is necessary to compare some numerical characteristics of the text, which would be approximate for the texts of the same author and differ significantly for the works of different authors. Such a characteristic can be the density of the distribution of letter combinations of three consecutive symbols (3-grams). During the experimental testing based on the developed four different algorithms for calculating the degree of verification of the author of the Ukrainian-language text from a set of possible values, values were obtained that confirm that the style of the authors numbered x and y is

quite close (more than 90%) to the style of collective works 1–4, respectively. Also, the number of authors (from 42.02% to 34.04% of the total 100 participants in the project from more than 300 articles) was significantly reduced, with similarity in speech style. Figure 13 presents graphs of the results obtained when applying algorithms to analyse the method developed to determine the author's style.

Further, an analysis of stop words and keywords of the authors' works was used to determine the author's style, as 34.04% got to those. Each individual has their vocabulary for conveying thought, including so-called "parasitic" (that is, therefore, although, etc.) and service words (and, and, and, but, although, etc.). Figure 14 presents an example of the analysis of the author's style in the second stage by analysing the frequency of service appearance and keywords, considering various filters. Therefore, a method of determining the style of the author of thematic Ukrainian-language textual content was developed based on the analysis of keywords, stable word combinations, N-grams, linguistics and stylometry, which made it possible to determine the stylistic contribution of each of the authors and increase the accuracy of attribution of a scientific and technical publication by 6%. A method for calculating the degree of verification of the author of a Ukrainian-language text from a set of possible ones based on a comparative analysis of the styles of potential authors has also been developed, which made it possible to increase the accuracy of classification by style similarity by 7%.

5. Conclusions

The work solves an important scientific and applied problem of analysis and synthesis of CLS for solving various problems of processing Ukrainian-language textual content based on the development of new and improvement of known models, methods and tools of NLP:

1. An analysis of the current state and prospects of IT development of natural language processing was carried out, which made it possible to define the problem and research tasks, as well as to form general research directions in the absence of non-commercial open-source software as CLS for processing Ukrainian-language textual content and a standardized design approach.
2. The relevance of solving the problem of analysis and synthesis of CLS based on the development of the general structure of the system for processing Ukrainian-language textual content is substantiated due to the interaction of the main processes/components of IS and methods of linguistic processing of textual content adapted to the Ukrainian language based on grapheme, morphological, lexical, syntactic, semantic, structural, ontological and pragmatic analysis allowed to improve the IT of intellectual analysis of text flow for solving a specific task of NLP. It ensured the adaptation of NLP processes for the analysis of Ukrainian-language textual content and, based on them, increased the accuracy of the obtained results by 6-48%, depending on the specific task of NLP. For example, for the NLP task of determining the Ukrainian-language text keywords, the density of keywords increases in the range [1.23; 1.48] times or by [23.14; 47.83]% depending on filling the thematic dictionary quality/accuracy through machine learning.
3. The methods of processing information resources, such as integration, management and support of Ukrainian-language content, were improved, which made it possible to adapt the process of intellectual analysis of the text flow and develop metrics of the effectiveness of the CLS functioning for the solution of various tasks of the NLP. The developed methods and tools make it possible to build a CLS for processing Ukrainian-language text content according to the needs of the permanent/potential target audience based on the analysis of the history of actions of website users.
4. The NLP methods based on regular expressions of pattern matching were improved, which made it possible to adapt the methods of tokenization and text normalization by cascades of simple substitutions of regular expressions and finite state machines.
5. The MA method of the Ukrainian-language text based on word segmentation and normalization, sentence segmentation and modified Porter's stemming algorithm was

- improved as an effective tool of identifying lemmata affixes for the possibility of marking the analysed word, which made it possible to increase the keyword searches accuracy by 9%.
6. The IT of the intellectual analysis of the text flow was improved based on the processing of information resources, which made it possible to adapt the general structure of modules for integration, management and support of content to solve various tasks of the NLP and increase the efficiency of the operation of the CLS by 6-9%. It became possible thanks to the combination of methods of linguistic analysis adapted to the Ukrainian language, improved IT processing of information resources, ML, and a set of metrics for evaluating the effectiveness of the CLS's functioning. The main principle of building such CLS is modularity, which facilitates their construction by requiring the availability of appropriate processes for solving a specific NLP problem.
 7. A method of determining the author in Ukrainian-language texts has been developed based on the analysis of the coefficients of the author's lexical speech in the reference passage of the author's text, which is based on the study of a collection of keywords, persistent phrases, indicators of linguometry, stylometry, as well as the results of the analysis of N-grams based on comparisons of usage differences 2-gram and 3-gram for publications similar in style in the range of [6;7]%, and for exactly not similar – >12%), which made it possible to determine a set of potential authors of publications from more than one author (up to [9; 34]% of the total number of project participants) and develop a method for identifying the author's style.
 8. A method of determining stable word combinations was developed based on the identification of keywords of the Ukrainian-language text and the analysis of the linguistic speech coefficients of the author of the text in reference excerpts of the content, which made it possible to improve the accuracy of the method of determining the style of the author of the text by 9% based on statistical linguistics.
 9. Relevant materials confirm the reliability of scientific and practical results on the implementation of dissertation studies by comparing the obtained practical results on different samples of reliable input data. CLS was developed using CMS Joomla on the information resource <http://victana.lviv.ua>! (for designing the e-framework of articles), PHP (for implementing text content processing methods), HTML (for implementing page markup), CSS (for describing page styles), and MySQL (for storing data and dictionaries). The experimental study confirmed the reliability of the method of identifying keywords - for different algorithms for processing the primary text, the average match between the lists of identified keywords and the author's keywords varies in the 52.6-68.5% range. The accuracy of matching keywords with the author's keywords ranges from 43.6 to 62.9%. The average match of meaningful keywords compared to all found by the system ranges from 38.9-75.8%, depending on the stages of analysis of article texts. The accuracy of matching keywords compared to all found by the system varies between 34.3-71.9%, depending on the stages of analysis of article texts.

Acknowledgements

The research was carried out with the grant support of the National Research Fund of Ukraine, "Information system development for automatic detection of misinformation sources and inauthentic behaviour of chat users ", project registration number 187/0012 from 1/08/2024 (2023.04/0012). Also, we would like to thank the reviewers for their precise and concise recommendations that improved the presentation of the results obtained.

References

- [1] I. Lauriola, A. Lavelli, F. Aioli, An introduction to deep learning in natural language processing: Models, techniques, and tools, *Neurocomputing* 470 (2022) 443-456.

- [2] Y. Kang, Z. Cai, C. W. Tan, Q. Huang, H. Liu, Natural language processing (NLP) in management research: A literature review, *Journal of Management Analytics* 7(2) (2020) 139-172.
- [3] L. Hickman, S. Thapa, L. Tay, M. Cao, P. Srinivasan, Text preprocessing for text mining in organizational research: Review and recommendations, *Organizational Research Methods* 25(1) (2022) 114-146.
- [4] D. Hu, An introductory survey on attention mechanisms in NLP problems, in: *Proceedings of the Intelligent Systems Conference on Intelligent Systems and Applications* 2 (2020) 432-448.
- [5] M. Gardner, W. Merrill, J. Dodge, M. E. Peters, A. Ross, S. Singh, N. A. Smith, Competency problems: On finding and removing artifacts in language data, *arXiv preprint arXiv:2104.08646*, 2021.
- [6] L. Wu, et. al., Graph neural networks for natural language processing: A survey, *Foundations and Trends in Machine Learning* 16(2) (2023) 119-328.
- [7] E. Fedorov, O. Nechyporenko, Linguistic Constructions Translation Method Based on Neural Networks, *CEUR Workshop Proceedings* 3396 (2023) 295-306.
- [8] M.-A. Lefer, N. Grabar, Super-creative and over bureaucratic: A cross-genre corpus based study on the use and translation of evaluative prefixation in ted talks and EU parliamentary debates, *Across Languages and Cultures* 16(2) (2015) 187–208.
- [9] M. Konyk, V. Vysotska, S. Goloshchuk, R. Holoshchuk, S. Chyrun, I. Budz, Technology of Ukrainian-English Machine Translation Based on Recursive Neural Network as LSTM, *CEUR Workshop Proceedings* 3387 (2023) 357-370.
- [10] N. Shakhovska, I. Shvorob, The method for detecting plagiarism in a collection of documents, in: *Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT*, 2015, pp. 142-145.
- [11] O. Karnalim, G. Kurniawati, Programming Style on Source Code Plagiarism and Collusion Detection, *International Journal of Computing* 19(1) (2020). 27-38.
- [12] V. Vysotska, Y. Burov, V. Lytvyn, A. Demchuk, Defining Author's Style for Plagiarism Detection in Academic Environment, in: *Proceedings of the International Conference on Data Stream Mining and Processing, DSMP*, 2018, pp. 128-133.
- [13] O. Barkovska, V. Kholiev, A. Havrashenko, D. Mohylevskyi, A. Kovalenko, A Conceptual Text Classification Model Based on Two-Factor Selection of Significant Words, *CEUR Workshop Proceedings* 3396 (2023) 244-255.
- [14] A. Berko, Y. Matseliukh, Y. Ivaniv, L. Chyrun, V. Schuchmann, The text classification based on Big Data analysis for keyword definition using stemming, in: *Proceedings of the IEEE 16th International conference on computer science and information technologies on Computer science and information technologies*, Lviv, Ukraine, 22–25 September, 2021, pp. 184–188.
- [15] V. Lytvyn, V. Vysotska, I. Budz, Y. Pelekh, N. Sokulska, R. Kovalchuk, L. Dzyubyk, O. Tereshchuk, M. Komar, Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution, *Eastern-European Journal of Enterprise Technologies*, 6(2(102)) (2019) 28–51. doi:10.15587/1729-4061.2019.186834.
- [16] I. Khomytska, I. Bazylevych, V. Teslyuk, I. Karamysheva, The chi-square test and data clustering combined for author identification, in: *Proceedings of the IEEE XVIIIth Scientific and Technical Conference on Computer Science and Information Technologies*, 2023.
- [17] I. Khomytska, V. Teslyuk, The Multifactor Method Applied for Authorship Attribution on the Phonological Level, *CEUR workshop proceedings* 2604 (2020) 189-198.
- [18] R. Romanchuk, V. Vysotska, V. Andrunyk, L. Chyrun, S. Chyrun, O. Brodyak, Intellectual Analysis System Project for Ukrainian-language Artistic Works to Determine the Text Authorship Attribution Probability, in: *Proceedings of the International Scientific and Technical Conference on Computer Sciences and Information Technologies*, 2023.
- [19] I. Khomytska, V. Teslyuk, A. Holovatyy, O. Morushko, Development of methods, models, and means for the author attribution of a text, *Eastern-European Journal of Enterprise Technologies* 3(2(93)) (2018) 41–46. doi: 10.15587/1729-4061.2018.132052.

- [20] I. Khomytska, V. Teslyuk, Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level, *Advances in Intelligent Systems and Computing* 871 (2019) 105–118. doi: 10.1007/978-3-030-01069-0_8.
- [21] R. Nazarchuk, S. Albota, Tweets about Ukraine during the russian-Ukrainian War: Quantitative Characteristics and Sentiment Analysis, *CEUR Workshop Proceedings* 3426 (2023) 551-560.
- [22] A. Taran, Terminology of Computational Linguistics in Terms of Indexing and Information Retrieval in the System "iSybislaw", *CEUR Workshop Proceedings* 2870 (2021) 225-234.
- [23] N. Kunanets, H. Matsiuk, Use of the Smart City Ontology for Relevant Information Retrieval, *CEUR Workshop Proceedings* 2362 (2019) 322-333.
- [24] K. Nataliia, M. Halyna, Application of Saaty Method While Choosing Thesaurus View Model of the "Smart city" Subject Domain for the Improvement of Information Retrieval Efficiency, in: *Proceedings of the IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT*, vol. 2, 2018, pp. 21-25. doi:10.1109/STC-CSIT.2018.8526656.
- [25] Y. Burov, V. Vysotska, L. Chyrun, Y. Ushenko, D. Uhryn, Z. Hu, Intelligent Network Architecture Development for E-Business Processes Based on Ontological Models, *International Journal of Information Engineering and Electronic Business* 16(5) (2024) 1-54. doi:10.5815/ijieeb.2024.05.01.
- [26] P. Zweigenbaum, S.J. Darmoni, N. Grabar, The contribution of morphological knowledge to French MeSH mapping for information retrieval, in: *Proceedings of the Annual AMIA Symposium*, 2001, pp. 796–800.
- [27] É. Bigeard, F. Thiessard, N. Grabar, Detecting drug non-compliance in internet fora using information retrieval and machine learning approaches, *Studies in Health Technology and Informatics* 264 (2019) 30–34.
- [28] V. Claveau, T. Hamon, S. Le Maguer, N. Grabar, Health consumer-oriented information retrieval, *Studies in Health Technology and Informatics* 210 (2015) 80–84.
- [29] V. Lytvyn, Y. Burov, V. Vysotska, Y. Pukach, O. Tereshchuk, I. Shakleina, Abstracting Text Content Based on Weighing the TF-IDF Measure by the Subject Area Ontology, in: *Proceedings of the IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Nur-Sultan, Kazakhstan, 2021. URL: <https://ieeexplore.ieee.org/document/9465978>.
- [30] A. Périnet, T. Hamon, Distributional analysis applied to specialized texts. Reduction of data sparseness by context abstractions, *Traitement Automatique des Langues* 56(2) (2015) 77–102.
- [31] V. Trysnyuk, Y. Nagornyi, K. Smetanin, I. Humeniuk, T. Uvarova, A method for user authenticating to critical infrastructure objects based on voice message identification, *Advanced Information Systems* 4(3) (2020) 11–16. doi:10.20998/2522-9052.2020.3.02.
- [32] O. Bisikalo, O. Boivan, N. Khairova, O. Kovtun, V. Kovtun, Precision automated phonetic analysis of speech signals for information technology of text-dependent authentication of a person by voice, *CEUR Workshop Proceedings* 2853 (2021) 276–288.
- [33] A. Sartiukova, O. Markiv, V. Vysotska, I. Shakleina, N. Sokulska, I. Romanets. Remote Voice Control of Computer Based on Convolutional Neural Network, in: *Proceedings of the IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Dortmund, Germany, 7 September 2023, pp. 1058-1064.
- [34] S. Kubinska, R. Holoshchuk, S. Holoshchuk, L. Chyrun, Ukrainian Language Chatbot for Sentiment Analysis and User Interests Recognition based on Data Mining, *CEUR Workshop Proceedings* 3171 (2022) 315-327.
- [35] V. Husak, O. Lozynska, I. Karpov, I. Peleshchak, S. Chyrun, A. Vysotskyi, Information System for Recommendation List Formation of Clothes Style Image Selection According to User's Needs Based on NLP and Chatbots, *CEUR Workshop Proceedings* 2604 (2020) 788-818.
- [36] A. Medvedyk, M. Lohoida, Z. Rybchak, O. Kulyna, IT Slang: Development of Telegram Chatbot, *CEUR Workshop Proceedings* 3396 (2023) 152-162.
- [37] O. Romanovskyi, N. Pidbutska, A. Knysh, Elomia Chatbot: The Effectiveness of Artificial Intelligence in the Fight for Mental Health, *CEUR Workshop Proceedings* 2870 (2021) 1215-1224.

- [38] A. Yarovy, D. Kudriavtsev, Method of Multi-Purpose Text Analysis Based on a Combination of Knowledge Bases for Intelligent Chatbot, CEUR Workshop Proceedings 2870 (2021) 1238-1248.
- [39] N. Shakhovska, O. Basystiuk, K. Shakhovska, Development of the Speech-to-Text Chatbot Interface Based on Google API, CEUR Workshop Proceedings 2386 (2019) 212-221.
- [40] T. Basyuk, A. Vasyliuk, Peculiarities of an Information System Development for Studying Ukrainian Language and Carrying out an Emotional and Content Analysis, CEUR Workshop Proceedings 3396 (2023). URL: <https://ceur-ws.org/Vol-3396/paper23.pdf>.
- [41] V. Vysotska, S. Holoshchuk, R. Holoshchuk, A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach, CEUR Workshop Proceedings 2870 (2021) 311-356.
- [42] A. Dmytriv, S. Holoshchuk, L. Chyrun, R. Holoshchuk, Comparative Analysis of Using Different Parts of Speech in the Ukrainian Texts Based on Stylistic Approach, CEUR Workshop Proceedings 3171 (2022) 546-560.
- [43] S. Yevseiev, et. al., Development of a Method for Determining the Indicators of Manipulation Based on Morphological Synthesis, Eastern-European Journal of Enterprise Technologies 117(9) (2022).
- [44] O. Cherednichenko, O. Kanishcheva, O. Yakovleva, D. Arkatov, Collection and Processing of a Medical Corpus in Ukrainian, CEUR Workshop Proceedings 2604 (2020) 272-282.
- [45] A. Dmytriv, V. Vysotska, M. Bublyk, The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using, in: Proceedings of the 16th International Conference on Computer Sciences and Information Technologies (CSIT), vol. 2, 2021, September, pp. 21-33.
- [46] V. Vysotska, S. Mazepa, L. Chyrun, O. Brodyak, I. Shakleina, V. Schuchmann, NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content, in: Proceedings of the IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, November, pp. 93-98.
- [47] M. Lupei, O. Mitsa, V. Sharkan, S. Vargha, N. Lupei, Analyzing Ukrainian Media Texts by Means of Support Vector Machines: Aspects of Language and Copyright, in: Proceedings of the International Conference on Computer Science, Engineering and Education Applications, 2023, March, pp. 173-182.
- [48] V. Vysotska, Analytical Method for Social Network User Profile Textual Content Monitoring Based on the Key Performance Indicators of the Web Page and Posts Analysis, CEUR Workshop Proceedings 3171 (2022) 1380-1402.
- [49] K. Shakhovska, I. Dumyn, N. Kryvinska, M. K. Kagita, An approach for a next-word prediction for Ukrainian language, Wireless Communications and Mobile Computing 2021 (2021) 1-9.
- [50] I. Demydov, Architecture of the Computer-linguistic System for Processing of Specialized Web-communities' Educational Content. URL: <https://ceur-ws.org/Vol-2616/paper1.pdf>.
- [51] V. Vysotska, Ukrainian participles formation by the generative grammars use, CEUR Workshop Proceedings 2604 (2020) 407-427.
- [52] B. Bengfort, R. Bilbro, T. Ojeda, Applied text analysis with Python: Enabling language-aware data products with machine learning, O'Reilly Media, Inc., 2018.
- [53] D. Jurafsky, J. H. Martin, Speech and Language Processing. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf.
- [54] D. Jurafsky, J. H. Martin, Regular Expressions, Text Normalization, Edit Distance. URL: <https://web.stanford.edu/~jurafsky/slp3/2.pdf>.
- [55] D. Jurafsky, J. H. Martin, Deep Learning Architectures for Sequence Processing. URL: <https://web.stanford.edu/~jurafsky/slp3/9.pdf>.
- [56] D. Jurafsky, J. H. Martin, Naive Bayes and Sentiment Classification. URL: <https://web.stanford.edu/~jurafsky/slp3/4.pdf>.
- [57] D. Jurafsky, J. H. Martin, Logistic Regression. URL: <https://web.stanford.edu/~jurafsky/slp3/5.pdf>.
- [58] D. Jurafsky, J. H. Martin, Neural Networks and Neural Language Models. URL: <https://web.stanford.edu/~jurafsky/slp3/7.pdf>.

- [59] I. Khomytska, V. Teslyuk, N. Kryvinska, I. Bazylevych, Software-based approach towards automated authorship acknowledgement-chi-square test on one consonant group, *Electronics (Switzerland)* 9(7) (2020) 1–11.
- [60] A. R. Sydor, V. M. Teslyuk, P. Y. Denysyuk, Recurrent expressions for reliability indicators of compound electropower systems, *Technical Electrodynamics* 4 (2014) 47–49.
- [61] V. Lytvyn, et. al., Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients, *Eastern-European Journal of Enterprise Technologies* 5(2(95)), (2018) 16–28. doi: 10.15587/1729-4061.2018.142451.
- [62] V. Lytvyn, et. al., Development of the system to integrate and generate content considering the cryptocurrent needs of users, *Eastern-European Journal of Enterprise Technologies* 1(2(97)) (2019) 18–39. doi: 10.15587/1729-4061.2019.154709.
- [63] P. Kravets, The Game Method for Orthonormal Systems Construction, in: *Proceedings of the 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics*, 2007. doi: 10.1109/cadsm.2007.4297555.